

***Q*- and *A*-learning Methods for Estimating Optimal Dynamic Treatment Regimes**

Phillip J. Schulte, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian¹

Abstract

In clinical practice, physicians make a series of treatment decisions over the course of a patient's disease based on his/her baseline and evolving characteristics. A dynamic treatment regime is a set of sequential decision rules that operationalizes this process. Each rule corresponds to a key decision point and dictates the next treatment action among the options available as a function of accrued information on the patient. Using data from a clinical trial or observational study, a key goal is estimating the optimal regime, that, if followed by the patient population, would yield the most favorable outcome on average. *Q*-learning and advantage (*A*-)learning are two main approaches for this purpose. We provide a detailed account of *Q*- and *A*-learning and study systematically the performance of these methods. The methods are illustrated using data from a study of depression.

1 Introduction

In the health sciences, an area of considerable current interest is personalized medicine, which involves making treatment decisions for an individual patient using all information available on the patient, including genetic, physiologic, demographic, and other clinical variables, to achieve the “best” outcome for the patient given this information. In treating a patient with an ongoing disease or disorder, a clinician makes a series of decisions based on the patient's evolving status, so seeking to tailor treatment to the patient. A dynamic treatment regime is a list of sequential decision rules that formalizes this process. Each rule corresponds to a key decision point in the disease or disorder progression and takes as input the information on the patient to that point and outputs the treatment that s/he should receive from among the available options. A key step toward personalized medicine is thus finding the optimal dynamic treatment regime, that which, if followed by the entire patient population, would yield the most favorable outcome on average.

¹Phillip J. Schulte is Graduate Student, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (E-mail pjschulte@ncsu.edu). Anastasios A. Tsiatis is Gertrude M. Cox Distinguished Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (E-mail tsiatis@ncsu.edu). Eric B. Laber is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (E-mail laber@stat.ncsu.edu). Marie Davidian is William Neal Reynolds Professor, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203, USA (E-mail davidian@ncsu.edu).

The statistical problem is to estimate the optimal regime based on data from a clinical trial or observational study. Q -learning (Q denoting “quality,” [Watkins, 1989](#); [Watkins and Dayan, 1992](#); [Nahum-Shani et al., 2010](#)) and advantage learning (A -learning, [Murphy, 2003](#); [Blatt, Murphy, and Zhu, 2004](#)) are two main approaches proposed for this purpose. Both follow from developments on reinforcement learning methods for sequential decision-making in the computer science literature. As described shortly, Q -learning is based roughly on posited regression models for the outcome of interest given patient information at each decision point and is implemented through a backwards (in time) recursive fitting procedure that is related to the dynamic programming algorithm ([Bather, 2000](#)), a standard approach for deducing optimal sequential decisions. A -learning involves the same recursive strategy, but, instead of requiring full regression relationships to be posited, requires only models for the part of the outcome regression involved in representing contrasts among treatments along with models for the probability of observed treatment assignment given patient information at each decision point. As discussed in the sequel, this feature may make A -learning more robust to model misspecification than Q -learning for consistent estimation of the optimal treatment regime.

Examples of the use of Q - and A -learning and related methods to deduce optimal strategies for treatment of substance abuse, psychiatric disorders, cancer, and HIV infection and for dose adjustment in response to evolving patient status are given by (e.g., [Rosthøj et al., 2006](#); [Murphy et al., 2007a,b](#); [Zhao, Kosorok, and Zeng, 2009](#); [Henderson, Ansell, and Alshibani, 2010](#)). Related work includes [Thall, Millikan, and Sung \(2000\)](#), [Thall, Sung, and Etsey \(2002\)](#), [Robins \(2004\)](#), [Moodie, Richardson, and Stephens \(2007\)](#), [Thall et al. \(2007\)](#), [Robins, Orellana, and Rotnitzky \(2008\)](#), [Almirall, Ten Have, and Murphy \(2010\)](#) and [Orellana, Rotnitzky, and Robins \(2010\)](#).

Despite increasing interest in estimation of optimal dynamic treatment regimes, there has been little study of the relative merits of Q - and A -learning, nor of consequences of misspecification of the postulated models involved. Moreover, although descriptions of Q - and A -learning are available, a self-contained account of both has not been presented. In this article, we provide a detailed description of an appropriate statistical framework in which an optimal regime may be defined formally and introduce Q - and A -learning in this context. Conditions under which these methods may be expected to yield credible estimators for

optimal regimes based on observed data are discussed, and we report on a systematic study of the methods' performance.

Section 2 introduces the statistical framework, and Section 3 makes precise the form of an optimal regime. We describe and contrast Q - and A -learning in Section 4 and present extensive simulations evaluating their performance, including under model misspecification, in Section 5. The methods are demonstrated using data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D, [Rush et al., 2004](#)) study in Section 6.

2 Framework and Assumptions

We consider the general setting of K prespecified, ordered decision points, indexed by $k = 1, \dots, K$, which may be times or events in the disease or disorder process that necessitate a treatment decision, where, at each point, a set of treatment options is available. Assume that there is a final outcome Y of interest for which, without loss of generality, large values are preferred. The outcome may be ascertained following the K th decision, as in the case of CD4 T-cell count at a prespecified follow-up time in a study of HIV infection ([Moodie et al., 2007](#)); or may be a function of information accrued over the entire sequence of decisions, as in [Henderson et al. \(2010\)](#), where outcome is the overall proportion of time a measure of blood clotting speed is kept within a target range in a study of dosing of anticoagulant agents.

In order to define an optimal treatment regime and discuss estimation of an optimal regime based on data from an observational study or clinical trial, we first define a suitable conceptual framework. For simplicity, our presentation is heuristic. We imagine that there is a superpopulation of patients, denoted by Ω , where one may view an element $\omega \in \Omega$ as a patient from this population. We assume that patients in the population have been treated and otherwise have behaved according to routine clinical practice for the disease or disorder prior to the first treatment decision. Consequently, immediately prior to this first decision, patient ω would present to the decision-maker with a set of baseline information (covariates) denoted by the random variable S_1 ; we discuss this further below. Thus, $S_1(\omega)$ is the value of his/her information immediately prior to decision 1 under these conditions, taking values s_1 , say, in a set \mathcal{S}_1 . Assume that, at each decision point $k = 1, \dots, K$, there is a set of possible treatment options \mathcal{A}_k , where

we denote elements of \mathcal{A}_k by a_k . We write $\bar{a}_k = (a_1, \dots, a_k)$ to denote a possible treatment history that could be administered through the k th decision, taking values in the corresponding set $\bar{\mathcal{A}}_k = \mathcal{A}_1 \times \dots \times \mathcal{A}_k$. Thus, $\bar{\mathcal{A}}_K$ denotes the set of all possible full treatment histories \bar{a}_K through all K decisions.

We then define the potential outcomes ([Robins, 1986](#))

$$W = \{S_2^*(a_1), S_3^*(\bar{a}_2), \dots, S_k^*(\bar{a}_{k-1}), \dots, S_K^*(\bar{a}_{K-1}), Y^*(\bar{a}_K) \text{ for all } \bar{a}_K \in \bar{\mathcal{A}}_K\}. \quad (1)$$

In (1), $S_k^*(\bar{a}_{k-1})(\omega)$ denotes the value of covariate information that would arise between the $(k-1)$ th and k th decision points for a patient $\omega \in \Omega$ under the hypothetical situation that s/he were to have received previously treatment history \bar{a}_{k-1} , taking values s_k in a set \mathcal{S}_k , $k = 2, \dots, K$. Similarly, $Y^*(\bar{a}_K)(\omega)$ is the hypothetical outcome that would result for patient ω were s/he to have been administered the full set of K treatments in \bar{a}_K . Here and henceforth, this notation implies that, for random variables such as $S_k^*(\bar{a}_{k-1})$, \bar{a}_{k-1} is an index representing prior treatment history. For convenience, write $\bar{S}_k^*(\bar{a}_{k-1}) = \{S_1, S_2^*(a_1), \dots, S_k^*(\bar{a}_{k-1})\}$, $k = 1, \dots, K$, where $\bar{S}_k^*(\bar{a}_{k-1})(\omega)$ takes values \bar{s}_k in $\bar{\mathcal{S}}_k = \mathcal{S}_1 \times \dots \times \mathcal{S}_k$; note that this definition includes the baseline covariate S_1 and is taken equal to S_1 when $k = 1$. In what follows, for simplicity, we take all random variables to be discrete, but the results we present hold more generally.

Let the random variables $A_1^{(P)}, \dots, A_K^{(P)}$ denote the treatments that would be assigned to patients in the population at decisions $1, \dots, K$ under routine clinical practice, so that $A_k^{(P)}(\omega)$ is the treatment in \mathcal{A}_k that patient ω would receive at decision k , taking values $a_k \in \mathcal{A}_k$. By routine clinical practice, we mean the conditions under which patients in the population and their providers would make treatment decisions acting as they see fit, emphasized by the superscript (P) (for “population”), to be distinguished from those of a clinical trial, discussed later. Thus, the $A_k^{(P)}$ characterize the mechanism by which treatments are assigned in the population if patients and clinicians are left to their own devices. Likewise, define the random variables $S_k^{(P)}$, $k = 2, \dots, K$, to be the covariate information that would be observed on patients in the population between decisions $k-1$ and k under the treatment assignments $A_k^{(P)}$, taking values $s_k \in \mathcal{S}_k$; let $Y^{(P)}$ be the corresponding observed outcome, taking values y in a set \mathcal{Y} ; and define $\bar{A}_k^{(P)} = (A_1^{(P)}, \dots, A_k^{(P)})$, taking values $\bar{a}_k \in \bar{\mathcal{A}}_k$. Henceforth, as is standard, we make the consistency assumption (e.g., [Robins, 1994](#)) that the covariates and outcomes that would be observed under these

conditions are those that potentially would be seen under the treatments actually received; that is, for patient $\omega \in \Omega$, $S_k^{(P)}(\omega) = S_k^*\{\bar{A}_{k-1}^{(P)}(\omega)\}(\omega)$, $k = 2, \dots, K$, and $Y^{(P)}(\omega) = Y^*\{\bar{A}_K^{(P)}(\omega)\}(\omega)$. We also make the stable unit treatment value assumption (Rubin, 1978), which ensures that a patient's covariates and outcome are unaffected by how treatments are allocated to her/him and other patients.

Under this conceptualization, probabilities for events in Ω are induced by random sampling from this population, as are all probability distributions of the potential data above and observed data that would be obtained from studies carried out in the population. The goal of Q - and A -learning is to estimate the optimal treatment regime based on data from an observational study or clinical trial carried out in a random sample from this population.

A dynamic treatment regime $d = (d_1, \dots, d_K)$ is a set of rules that dictates an algorithm for treating a patient over time. At the k th decision point, the k th rule $d_k(\bar{s}_k, \bar{a}_{k-1})$, say, takes as input the patient's realized covariate and treatment history prior to the k th treatment decision and outputs a value $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1}) \subseteq \mathcal{A}_k$; for $k = 1$, there is no prior treatment (a_0 is null), and we write $d_1(s_1)$ and $\Psi_1(s_1)$. Here, $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$ is the set of feasible treatment options for a patient with realized history $(\bar{s}_k, \bar{a}_{k-1})$, reflecting that some treatment options may be unethical or impossible for patients with certain histories. We discuss considerations for identifying the $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$ shortly. Because we consider only regimes where $d_k(\bar{s}_k, \bar{a}_{k-1}) \in \Psi_k(\bar{s}_k, \bar{a}_{k-1}) \subseteq \mathcal{A}_k$, d_k need only map a subset of $\bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_{k-1}$ to \mathcal{A}_k . We define these subsets recursively as

$$\Gamma_k = \left\{ (\bar{s}_k, \bar{a}_{k-1}) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_{k-1} \text{ satisfying} \right. \\ \left. \text{(i) } a_j \in \Psi_j(\bar{s}_j, \bar{a}_{j-1}), j = 1, \dots, k-1, \text{ and } \text{(ii) } \text{pr}\{\bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k\} > 0 \right\} \quad (2)$$

for $k = 1, \dots, K$. Thus, we may define formally the class of all feasible treatment regimes \mathcal{D} , say, as the set of all $d = (d_1, \dots, d_K)$ for which d_k , $k = 1, \dots, K$, is a mapping from Γ_k into \mathcal{A}_k satisfying $d_k(\bar{s}_k, \bar{a}_{k-1}) \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$ for every $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$.

Intuitively, an optimal regime should represent the “best” way to intervene to treat patients in Ω who would otherwise behave according to routine clinical practice. We now state with specificity what we mean by this. To this end, for any $d \in \mathcal{D}$, writing $\bar{d}_k = (d_1, \dots, d_k)$, $k = 1, \dots, K$, $\bar{d}_K = d$, define the potential

outcomes $\{S_2^*(d_1), \dots, S_k^*(\bar{d}_{k-1}), \dots, S_K^*(\bar{d}_{K-1}), Y^*(d)\}$ associated with a regime $d \in \mathcal{D}$ such that, for any $\omega \in \Omega$, with $S_1(\omega) = s_1$,

$$\begin{aligned} d_1(s_1) = u_1, S_2^*(d_1)(\omega) = S_2^*(u_1)(\omega) = s_2, d_2(\bar{s}_2, u_1) = u_2, \dots, d_{K-1}(\bar{s}_{K-1}, \bar{u}_{K-2}) = u_{K-1}, \\ S_K^*(\bar{d}_{K-1})(\omega) = S_K^*(\bar{u}_{K-1})(\omega) = s_K, d_K(\bar{s}_K, \bar{u}_{K-1}) = u_K, Y^*(d)(\omega) = Y^*(\bar{u}_K)(\omega) = y. \end{aligned} \quad (3)$$

The index \bar{d}_{k-1} emphasizes that $\bar{S}_k^*(\bar{d}_{k-1})(\omega)$ represents the covariate information that would arise between decisions $k-1$ and k were patient ω to receive the treatments sequentially dictated by the first $k-1$ rules in d . Similarly, $Y^*(d)(\omega)$ is the final outcome that ω would experience if s/he were to receive the K treatments dictated by d .

With these definitions, the expected outcome in the population if all patients with initial state $S_1 = s_1$ were to follow regime d is $E\{Y^*(d)|S_1 = s_1\}$. An optimal regime, $d^{\text{opt}} \in \mathcal{D}$, say, satisfies

$$E\{Y^*(d)|S_1 = s_1\} \leq E\{Y^*(d^{\text{opt}})|S_1 = s_1\} \quad \text{for all } d \in \mathcal{D} \text{ and all } s_1 \in \mathcal{S}_1. \quad (4)$$

In Section 3, we give the form of d^{opt} satisfying (4) and demonstrate further optimality properties.

Of course, potential outcomes for a given patient for all $d \in \mathcal{D}$ are not observed. Thus, the goal is to estimate d^{opt} in (4) using data from a study carried out on a random sample of n patients from Ω that record baseline and evolving covariate information and the treatments actually received by the participants. We denote the available study data as independent and identically distributed (i.i.d.) time-ordered random variables $(S_{1i}, A_{1i}, \dots, S_{Ki}, A_{Ki}, Y_i)$ $i = 1, \dots, n$ on Ω . Here, S_1 is as before; S_k , $k = 2, \dots, K$, is the covariate information recorded between decisions $k-1$ and k , taking values $s_k \in \mathcal{S}_k$; A_k , $k = 1, \dots, K$, is the recorded, observed treatment assignment, taking values $a_k \in \mathcal{A}_k$; and Y is the observed outcome, taking values $y \in \mathcal{Y}$. As above, we define $\bar{S}_k = (S_1, \dots, S_k)$ and $\bar{A}_k = (A_1, \dots, A_k)$, $k = 1, \dots, K$, taking values $\bar{s}_k \in \bar{\mathcal{S}}_k$ and $\bar{a}_k \in \bar{\mathcal{A}}_k$.

It is important to recognize that the nature of the study generating the available data must be considered carefully. If the data arise from an observational study in which covariate, treatment, and outcome information on n participants randomly sampled from Ω is recorded, with no intervention by investiga-

tors, then it is reasonable to assume that the mechanism by which treatments are assigned to the patients in the sample during the study is the same as that for the entire population under routine practice. In this case, for $k = 1, \dots, K$, $A_k = A_k^{(P)}$, so that, under the consistency assumption, for $k = 2, \dots, K$, $S_k(\omega) = S_k^{(P)}(\omega) = S_k^*\{\bar{A}_{k-1}^{(P)}(\omega)\}(\omega)$, and $Y(\omega) = Y^{(P)}(\omega) = Y^*\{\bar{A}_K^{(P)}(\omega)\}(\omega)$. Here, the form of $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, \dots, K$, is determined by treatment choices dictated by clinical practice.

Such a correspondence between the S_k, A_k and $S_k^{(P)}, A_k^{(P)}$ is not the case for an intervention study. A clinical trial design that has been advocated for collecting data suitable for estimating optimal treatment regimes is that of a so-called sequential multiple-assignment randomized trial (SMART, [Lavori and Dawson, 2000](#); [Murphy, 2005](#)). In a SMART involving K pre-specified decision points, each participant is randomized at each decision point to one of a set of feasible treatment options, where, at the k th decision, the randomization probabilities may depend on past realized information \bar{s}_k, \bar{a}_{k-1} . As we discuss further shortly, as with any clinical trial, an advantage is that the usual issues of confounding associated with an observational study are obviated. However, the treatment assignment mechanism in the study is no longer the same as that in the population under routine practice. More precisely, the sample space is now $\Omega \times \bar{A}_K$, where for any element $(\omega \times \bar{a}_K)$, ω represents the patient randomly sampled from the population, and \bar{a}_K represents the treatments assigned to her/him at all K decisions by the random mechanism dictated by the trial design. Here, then, the observed $A_k(\omega \times \bar{a}_K) = a_k$ and $S_k(\omega \times \bar{a}_k) = S_k^*(\bar{a}_{k-1})(\omega)$. Thus, in contrast to an observational study, $A_k \neq A_k^{(P)}$ and $S_k \neq S_k^{(P)}$ in general. Moreover, the treatment options in $\Psi_k(\bar{s}_k, \bar{a}_{k-1})$ are dictated by the trial design so may be different from those in routine practice. In particular, the set of treatment options at each decision might be restricted relative to those available in clinical practice for reasons of logistics, cost, or interest of the trial sponsor in only certain products. We discuss further considerations for using data from a SMART to estimate optimal regimes in [Section A.2](#) of the Appendix.

In order to use the observed data from either type of study to estimate an optimal regime, the critical assumption of no unmeasured confounders, also referred to as the sequential randomization assumption ([Robins, 1994](#)), must be satisfied. A version of this assumption states that A_k is conditionally independent of W given $\{\bar{S}_k, \bar{A}_{k-1}\}$, $k = 1, \dots, K$, where A_0 is null, written $A_k \perp\!\!\!\perp W | \bar{S}_k, \bar{A}_{k-1}$. In a SMART, this

assumption is automatically satisfied by design. In an observational study, this assumption is unverifiable from the observed data. Although in the population patients and their providers may make treatment decisions based on past covariate information available to them, the issue is whether or not all of this information is recorded in the S_k ; see Section A.2 of the Appendix.

3 Defining the Optimal Treatment Regime

Q - and A -learning are two approaches to estimating d^{opt} satisfying (4) under the foregoing framework and assumptions. Both involve similar recursive fitting algorithms; the main distinguishing feature is the form of the respective underlying models. To appreciate the rationale for the methods, one must first understand how d^{opt} is determined via dynamic programming, also referred to as backward induction. We demonstrate the formulation of d^{opt} in terms of the potential outcomes and then show how d^{opt} may be expressed in terms of the observed data under assumptions including those of the last section. In the following, we sometimes highlight dependence on specific elements of quantities such as \bar{a}_k , writing, for example, \bar{a}_k as (\bar{a}_{k-1}, a_k) .

At the K th decision point, for any $\bar{s}_K \in \bar{\mathcal{S}}_K$, $\bar{a}_{K-1} \in \bar{\mathcal{A}}_{K-1}$ for which $(\bar{s}_K, \bar{a}_{K-1}) \in \Gamma_K$, define

$$d_K^{(1)\text{opt}}(\bar{s}_K, \bar{a}_{K-1}) = \arg \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\{Y^*(\bar{a}_{K-1}, a_K) | \bar{S}_K^*(\bar{a}_{K-1}) = \bar{s}_K\}, \quad (5)$$

$$V_K^{(1)}(\bar{s}_K, \bar{a}_{K-1}) = \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\{Y^*(\bar{a}_{K-1}, a_K) | \bar{S}_K^*(\bar{a}_{K-1}) = \bar{s}_K\}. \quad (6)$$

For $k = K-1, \dots, 1$ and any $\bar{s}_k \in \bar{\mathcal{S}}_k$, $\bar{a}_{k-1} \in \bar{\mathcal{A}}_{k-1}$ for which $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$, let

$$d_k^{(1)\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = \arg \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} \mathbb{E}[V_{k+1}^{(1)}\{\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\} | \bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k], \quad (7)$$

$$V_k^{(1)}(\bar{s}_k, \bar{a}_{k-1}) = \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} \mathbb{E}[V_{k+1}^{(1)}\{\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\} | \bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k], \quad (8)$$

so that, for $s_1 \in \mathcal{S}_1$, $d_1^{(1)\text{opt}}(s_1) = \arg \max_{a_1 \in \Psi_1(s_1)} \mathbb{E}[V_2^{(1)}\{s_1, S_2^*(a_1), a_1\} | S_1 = s_1]$, and $V_1^{(1)}(s_1) = \max_{a_1 \in \Psi_1(s_1)} \mathbb{E}[V_2^{(1)}\{s_1, S_2^*(a_1), a_1\} | S_1 = s_1]$. Note that the above conditional expectations are well-defined by condition (ii) in (2) defining Γ_k .

It is clear that $d^{(1)\text{opt}} = (d_1^{(1)\text{opt}}, \dots, d_K^{(1)\text{opt}})$ defined above is a treatment regime, as it comprises a set of rules that uses patient information prior to each decision to assign treatment from among the feasible options. The superscript (1) indicates that $d^{(1)\text{opt}}$ provides a set of K rules for a patient presenting prior to decision point 1 with baseline information $S_1 = s_1$. Note that $d^{(1)\text{opt}}$ is defined in a backward iterative fashion. At the K th decision, (5) gives the treatment among the feasible options at decision K that maximizes the expected potential final outcome given the prior potential information available, and (6) is the maximum value achieved. At decisions $k = K - 1, \dots, 1$, intuitively, (7) gives the treatment that maximizes the expected outcome that would be achieved if subsequent optimal rules already defined were followed henceforth.

In Section A.1 of the Appendix, we provide a formal argument demonstrating that $d^{(1)\text{opt}}$ defined in (5)–(8) is an optimal treatment regime in the sense of satisfying (4). Note that, because (4) is true for any s_1 , in fact $E\{Y^*(d)\} \leq E\{Y^*(d^{(1)\text{opt}})\}$ for any $d \in \mathcal{D}$. Thus, from a policy perspective, $d^{(1)\text{opt}}$ defines the optimal strategy for treating patients in the population through all K decisions were they to be encountered at the stage of the disease or disorder that precedes decision point 1.

In routine clinical practice, however, patients may be encountered at later stages. Consider a patient $\omega \in \Omega$ for whom the first $\ell - 1$ treatment decisions have been made as seen fit by her/him and her/his provider, $\ell = 2, \dots, K$. Immediately prior to the ℓ th decision, the patient would have past history $\bar{S}_\ell^{(P)}(\omega) = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)}(\omega) = \bar{a}_{\ell-1}$, raising the issue of how best to intervene to treat such a patient henceforth, from the ℓ th to K th decisions. That is, we desire rules $d_k^{(\ell)}(\bar{s}_k, \bar{a}_{k-1})$, $k = \ell, \ell + 1, \dots, K$, say, that dictate how to treat such patients.

Write $d^{(\ell)} = (d_\ell^{(\ell)}, d_{\ell+1}^{(\ell)}, \dots, d_K^{(\ell)})$ to denote regimes starting at the ℓ th decision point. Analogous to the above, we define the class $\mathcal{D}^{(\ell)}$ of all such feasible regimes to be the set of all $d^{(\ell)}$ for which $d_k^{(\ell)}(\bar{s}_k, \bar{a}_{k-1}) = a_k$ for $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k^{(\ell)}$ and $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$ for $k = \ell, \dots, K$, where

$$\Gamma_k^{(\ell)} = \left\{ (\bar{s}_k, \bar{a}_{k-1}) \in \bar{\mathcal{S}}_k \times \bar{\mathcal{A}}_{k-1} \text{ satisfying} \right. \\ \left. \text{(i) } a_j \in \Psi_j(\bar{s}_j, \bar{a}_{j-1}), \ j = \ell, \dots, k-1, \text{ and (ii) } \text{pr}\{\mathcal{V}_{\ell,k}\} > 0 \right\},$$

$$\mathcal{V}_{\ell,k} \text{ is the event } \{\bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}, S_{\ell+1}^*(\bar{a}_\ell) = s_{\ell+1}, \dots, S_k^*(\bar{a}_{k-1}) = s_k\}.$$

Then, by analogy to (4), we seek $d^{(\ell)\text{opt}}$ satisfying

$$\mathbb{E}\{Y^*(\bar{a}_{\ell-1}, d^{(\ell)}) | \bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}\} \leq \mathbb{E}\{Y^*(\bar{a}_{\ell-1}, d^{(\ell)\text{opt}}) | \bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}\} \quad (9)$$

for all $d^{(\ell)} \in \mathcal{D}^{(\ell)}$ and $\bar{s}_\ell \in \bar{\mathcal{S}}_\ell$, $\bar{a}_{\ell-1} \in \bar{\mathcal{A}}_{\ell-1}$ for which $\text{pr}(\bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}) > 0$. Viewing this as a problem of making $K - \ell + 1$ decisions at decision points $\ell, \ell + 1, \dots, K$, with initial state $\bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}$, by an argument analogous to that in Section A.1 of the Appendix for $\ell = 1$ and initial state $S_1 = s_1$, it may be shown that $d^{(\ell)\text{opt}}$ satisfying (9) is given by

$$d_K^{(\ell)\text{opt}}(\bar{s}_K, \bar{a}_{K-1}) = \arg \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\{Y^*(\bar{a}_{K-1}, a_K) | \mathcal{V}_{\ell, K}\}, \quad (10)$$

$$V_K^{(\ell)}(\bar{s}_K, \bar{a}_{K-1}) = \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} \mathbb{E}\{Y^*(\bar{a}_{K-1}, a_K) | \mathcal{V}_{\ell, K}\} \quad (11)$$

for any $\bar{s}_K \in \bar{\mathcal{S}}_K$, $\bar{a}_{K-1} \in \bar{\mathcal{A}}_{K-1}$ for which $(\bar{s}_K, \bar{a}_{K-1}) \in \Gamma_K^{(\ell)}$; and, for $k = K - 1, \dots, \ell$,

$$d_k^{(\ell)\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = \arg \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} \mathbb{E}[V_{k+1}^{(\ell)}\{\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\} | \mathcal{V}_{\ell, k}], \quad (12)$$

$$V_k^{(\ell)}(\bar{s}_k, \bar{a}_{k-1}) = \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} \mathbb{E}[V_{k+1}^{(\ell)}\{\bar{s}_k, S_{k+1}^*(\bar{a}_{k-1}, a_k), \bar{a}_{k-1}, a_k\} | \mathcal{V}_{\ell, k}] \quad (13)$$

for any $\bar{s}_k \in \bar{\mathcal{S}}_k$, $\bar{a}_{k-1} \in \bar{\mathcal{A}}_{k-1}$ for which $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k^{(\ell)}$, so that

$$d_\ell^{(\ell)\text{opt}}(\bar{s}_\ell, \bar{a}_{\ell-1}) = \arg \max_{a_\ell \in \Psi_\ell(\bar{s}_\ell, \bar{a}_{\ell-1})} \mathbb{E}[V_{\ell+1}^{(\ell)}\{\bar{s}_\ell, S_{\ell+1}^*(\bar{a}_{\ell-1}, a_\ell), \bar{a}_{\ell-1}, a_\ell\} | \bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}].$$

Comparison of (5)–(8) to (10)–(13) shows that the ℓ th to K th rules of the optimal regime $d^{(1)\text{opt}}$ that would be followed by a patient presenting at the first decision are not necessarily the same as those of the optimal regime $d^{(\ell)\text{opt}}$ that would be followed by a patient presenting at the ℓ th decision. In particular, noting that the conditioning sets in (5)–(8) are $\mathcal{V}_{1, K}$ and $\mathcal{V}_{1, k}$, the rules are ℓ -dependent through dependence of the conditioning sets $\mathcal{V}_{\ell, k}$, $\ell = 1, \dots, K$, $k = \ell, \dots, K$, on ℓ . However, we demonstrate shortly that these rules coincide under certain conditions.

The foregoing developments define optimal regimes in terms of potential outcomes. To be useful in

practice, an optimal regime must be defined in terms of the observed data. To this end, define

$$Q_K(\bar{s}_K, \bar{a}_K) = E(Y|\bar{S}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K), \quad (14)$$

$$d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1}) = \arg \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} Q_K(\bar{s}_K, \bar{a}_{K-1}, a_K), \quad (15)$$

$$V_K(\bar{s}_K, \bar{a}_{K-1}) = \max_{a_K \in \Psi_K(\bar{s}_K, \bar{a}_{K-1})} Q_K(\bar{s}_K, \bar{a}_{K-1}, a_K), \quad (16)$$

for any $\bar{s}_K \in \bar{\mathcal{S}}_K, \bar{a}_K \in \bar{\mathcal{A}}_K$ for which $\text{pr}(\bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1}) > 0$; and for $k = K-1, \dots, 1$,

$$Q_k(\bar{s}_k, \bar{a}_k) = E\{V_{k+1}(\bar{s}_k, S_{k+1}, \bar{a}_k) | \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k\} \quad (17)$$

$$d_k^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = \arg \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k), \quad (18)$$

$$V_k(\bar{s}_k, \bar{a}_{k-1}) = \max_{a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})} Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k), \quad (19)$$

for any $\bar{s}_k \in \bar{\mathcal{S}}_k, \bar{a}_k \in \bar{\mathcal{A}}_k$ for which $\text{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) > 0$. Note that all quantities in (14)–(19) are expressed entirely in terms of the distribution of the observed data.

In Section A.2 of the Appendix, under the consistency and sequential randomization (no unmeasured confounders) assumptions, along with positivity assumptions on probabilities associated with events involving \bar{S}_K, \bar{A}_K and $\bar{S}_K^{(P)}, \bar{A}_K^{(P)}$ given in Section A.2 of the Appendix, we show that

$$\Gamma_k^{(\ell)} = \Gamma_k, \quad (20)$$

$$d_k^{(\ell)\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = d_k^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}), \quad (21)$$

$$V_k^{(\ell)}(\bar{s}_k, \bar{a}_{k-1}) = V_k(\bar{s}_k, \bar{a}_{k-1}), \quad (22)$$

for $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$ for $\ell = 1, \dots, K$ and $k = \ell, \dots, K$. The equivalence in (20)–(22) not only demonstrates that an optimal treatment regime can be obtained using the distribution of the observed data but also that the corresponding rules dictating treatment do not depend on ℓ under these assumptions. Thus, (20)–(22) imply that the single set of rules $d^{\text{opt}} = (d_1^{\text{opt}}, \dots, d_K^{\text{opt}})$ defined in (15) and (18) is relevant regardless of when a patient presents. That is, treatment at the ℓ th decision point for a patient who presents at decision 1 and has followed the rules in d^{opt} to that point would be determined by d_ℓ^{opt} evaluated at his/her history

up to that point, as would treatment for a subject presenting for the first time immediately prior to decision ℓ .

The $Q_k(\bar{s}_k, \bar{a}_k)$ in (14) and (17) are referred to as the “ Q -functions,” viewed as measuring the “quality” associated with using treatment a_k at decision k given the history up to that decision and then following the optimal regime thereafter. The “value functions” $V_k(\bar{s}_k, \bar{a}_{k-1})$ in (16) and (19) reflect the “value” of a patient’s history \bar{s}_k, \bar{a}_{k-1} assuming that optimal decisions are made in the future.

It is worth noting that there may not be a unique d^{opt} . At any decision point k , if there is more than one feasible treatment option a_k leading to the maximum value of the Q -function, then any rule d_k^{opt} yielding one of these a_k defines an optimal regime.

4 Q - and A -Learning

4.1 Q -Learning

From (15) and (18), the optimal regime d^{opt} is defined in terms of the Q -functions (14), (17). Thus, estimation of d^{opt} based on i.i.d. data $(S_{1i}, A_{1i}, \dots, S_{Ki}, A_{Ki}, Y_i)$, $i = 1, \dots, n$, may be accomplished via direct modeling and fitting of the Q -functions. This is the approach underlying Q -learning. Specifically, one may posit models $Q_k(\bar{s}_k, \bar{a}_k; \xi_k)$, say, for $k = K, K-1, \dots, 1$, each depending on a finite-dimensional parameter ξ_k . The models may be linear or nonlinear in ξ_k and include main effects and interactions in the elements of \bar{s}_k and \bar{a}_k .

Estimators $\hat{\xi}_k$ may be obtained in a backward iterative fashion for $k = K, K-1, \dots, 1$ by solving suitable estimating equations [e.g., ordinary (OLS) or weighted (WLS) least squares]. Assuming the latter, for $k = K$, letting $\tilde{V}_{(K+1)i} = Y_i$, one would first solve

$$\sum_{i=1}^n \frac{\partial Q_K(\bar{S}_{Ki}, \bar{A}_{Ki}; \xi_K)}{\partial \xi_K} \Sigma_K^{-1}(\bar{S}_{Ki}, \bar{A}_{Ki}) \{ \tilde{V}_{(K+1)i} - Q_K(\bar{S}_{Ki}, \bar{A}_{Ki}; \xi_K) \} = 0 \quad (23)$$

in ξ_K to obtain $\hat{\xi}_K$, where $\Sigma_K(\bar{s}_K, \bar{a}_K)$ is a working variance model. Substituting the model $Q_K(\bar{s}_K, \bar{a}_K; \xi_K)$ in (15) and accordingly writing $d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1}; \xi_K)$, substituting $\hat{\xi}_K$ for ξ_K yields an estimator for the

optimal treatment choice at decision K for a patient with past history $\bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1}$. With $\hat{\xi}_K$ in hand, one would form for each i , based on (16), $\tilde{V}_{Ki} = \max_{a_K \in \Psi_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i})} Q_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i}, a_K; \hat{\xi}_K)$. To obtain $\hat{\xi}_{K-1}$, setting $k = K - 1$, based on (17), one would then solve in ξ_k

$$\sum_{i=1}^n \frac{\partial Q_k(\bar{S}_{ki}, \bar{A}_{ki}; \xi_k)}{\partial \xi_k} \Sigma_k^{-1}(\bar{S}_{ki}, \bar{A}_{ki}) \{ \tilde{V}_{(k+1)i} - Q_k(\bar{S}_{ki}, \bar{A}_{ki}; \xi_k) \} = 0, \quad (24)$$

where $\Sigma_k(\bar{s}_k, \bar{a}_k)$ is a working variance model. The corresponding $d_{K-1}^{\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2}; \hat{\xi}_{K-1})$ then yields an estimator for the optimal treatment choice at decision $K - 1$ for a patient with past history $\bar{S}_{K-1} = \bar{s}_{K-1}, \bar{A}_{K-2} = \bar{a}_{K-2}$, assuming s/he will take the optimal treatment at decision K . One would continue this process in the obvious fashion for $k = K-2, \dots, 1$, forming $\tilde{V}_{ki} = \max_{a_k \in \Psi_k(\bar{S}_{ki}, \bar{A}_{(k-1)i})} Q_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}, a_k; \hat{\xi}_k)$, and solving equations of form (24) to obtain $\hat{\xi}_k$ and corresponding $d_k^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}; \hat{\xi}_k)$.

We may now summarize the estimated optimal regime as $\hat{d}_Q^{\text{opt}} = (\hat{d}_{Q,1}^{\text{opt}}, \dots, \hat{d}_{Q,K}^{\text{opt}})$, where

$$\hat{d}_{Q,1}^{\text{opt}}(s_1) = d_1^{\text{opt}}(s_1; \hat{\xi}_1), \dots, \hat{d}_{Q,k}^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = d_k^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}; \hat{\xi}_k), \quad k = 2, \dots, K. \quad (25)$$

It is important to recognize that the estimated regime (25) may not be a credible estimator for the true optimal regime unless all the models for the Q -functions are correctly specified.

We illustrate for the case $K = 2$, where at each decision there are two feasible treatment options coded as 0 and 1; i.e., $\Psi_1(s_1) = \mathcal{A}_1 = \{0, 1\}$ for all s_1 and $\Psi_2(\bar{s}_2, a_1) = \mathcal{A}_2 = \{0, 1\}$ for all \bar{s}_2 and $a_1 \in \{0, 1\}$. Let $\mathcal{H}_1 = (1, s_1^T)^T$ and $\mathcal{H}_2 = (1, s_1^T, a_1, s_2^T)^T$. As in many modeling contexts, it is standard to adopt linear models for the Q -functions; accordingly, consider the models

$$Q_1(s_1, a_1; \xi_1) = \mathcal{H}_1^T \beta_1 + a_1(\mathcal{H}_1^T \psi_1), \quad Q_2(\bar{s}_2, \bar{a}_2; \xi_2) = \mathcal{H}_2^T \beta_2 + a_2(\mathcal{H}_2^T \psi_2), \quad (26)$$

where $\xi_k = (\beta_k^T, \psi_k^T)^T$ for $k = 1, 2$. Note that $Q_2(\bar{s}_2, \bar{a}_2; \xi_2)$ in (26) is a model for $E(Y|\bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2)$, which is a standard regression problem involving observable data, whereas $Q_1(s_1, a_1; \xi_1)$ is a model for the conditional expectation of $V_2(\bar{s}_2, a_1) = \max_{a_2 \in \{0,1\}} E(Y|\bar{S}_2 = s_2, A_1 = a_1, A_2 = a_2)$ given $S_1 = s_1$ and $A_1 = a_1$, which is an approximation to a complex relationship involving a maximization. Under (26), it is

straightforward to deduce that $V_2(\bar{s}_2, a_1; \xi_2) = \max_{a_2 \in \{0,1\}} Q_2(\bar{s}_2, a_1, a_2; \xi_2) = \mathcal{H}_2^T \beta_2 + (\mathcal{H}_2^T \psi_2) I(\mathcal{H}_2^T \psi_2 > 0)$ and $V_1(s_1; \xi_1) = \max_{a_1 \in \{0,1\}} Q_1(s_1, a_1; \xi_1) = \mathcal{H}_1^T \beta_1 + (\mathcal{H}_1^T \psi_1) I(\mathcal{H}_1^T \psi_1 > 0)$. Substituting the Q -functions in (26) in (15) and (18) then yields $d_1^{\text{opt}}(s_1; \xi_1) = I(\mathcal{H}_1^T \psi_1 > 0)$ and $d_2^{\text{opt}}(\bar{s}_2, a_1; \xi_2) = I(\mathcal{H}_2^T \psi_2 > 0)$.

We have presented (23) and (24) in the conventional WLS form, with leading term in the summand $\partial/\partial \xi_k Q_k(\bar{S}_{ki}, \bar{A}_{ki}; \xi_k) \Sigma_k^{-1}(\bar{S}_{ki}, \bar{A}_{ki})$; taking Σ_k to be a constant yields OLS. At the K th decision, with responses Y_i , standard theory implies that this is the optimal leading term when $\text{var}(Y|\bar{S}_K = s_K, \bar{A}_K = a_K) = \Sigma_K(\bar{s}_K, \bar{a}_K)$, yielding the efficient (asymptotically) estimator for ξ_K . For $k < K$, with “responses” $\tilde{V}_{(k+1)i}$, this theory may no longer apply; however, deriving the optimal leading term involves considerable complication. Accordingly, it is standard to fit the posited models $Q_k(\bar{s}_k, \bar{a}_k; \xi_k)$ via OLS or WLS; some authors define Q -learning as using OLS (Chakraborty, Murphy, and Strecher, 2010). The choice may be dictated by apparent relevance of the homoscedasticity assumption on the $\tilde{V}_{(k+1)i}$, $k = K, K-1, \dots, 1$, and whether or not linear models are sufficient to approximate the relationships may also be evaluated, but see Section 4.3.

4.2 A-Learning

Advantage learning (A -learning, Murphy, 2003) is an alternative to Q -learning that involves making fewer assumptions on the form of the Q -functions. For simplicity, we consider the case of two feasible treatment options coded as 0 and 1 at each decision; i.e., $\Psi_k(\bar{s}_k, \bar{a}_{k-1}) = \mathcal{A}_k = \{0, 1\}$, $k = 1, \dots, K$, though the methodology can be extended to an arbitrary number of treatments at each stage at the expense of complicating the formulation and notation.

To fix ideas, consider (26). Note that $d_1^{\text{opt}}(s_1; \xi_1)$ implied by (26) depends only on $\mathcal{H}_1^T \psi_1 = Q_1(s_1, 1; \xi_1) - Q_1(s_1, 0; \xi_1)$; likewise, $d_2^{\text{opt}}(\bar{s}_2, a_1; \xi_2)$ depends on $\mathcal{H}_2^T \psi_2 = Q_2(\bar{s}_2, a_1, 1; \xi_2) - Q_2(\bar{s}_2, a_1, 0; \xi_2)$. This is a special case of the general result that, for purposes of deducing the optimal regime, for each $k = 1, \dots, K$, it suffices to know the contrast function $C_k(\bar{s}_k, \bar{a}_{k-1}) = Q_k(\bar{s}_k, \bar{a}_{k-1}, 1) - Q_k(\bar{s}_k, \bar{a}_{k-1}, 0)$. This can be appreciated by noting that any arbitrary $Q_k(\bar{s}_k, \bar{a}_k)$ may be written as $h_k(\bar{s}_k, \bar{a}_{k-1}) + a_k C_k(\bar{s}_k, \bar{a}_{k-1})$, where $h_k(\bar{s}_k, \bar{a}_{k-1}) = Q_k(\bar{s}_k, \bar{a}_{k-1}, 0)$, so that $Q_k(\bar{s}_k, \bar{a}_{k-1}, a_k)$ is maximized by taking $a_k = I\{C_k(\bar{s}_k, \bar{a}_{k-1}) > 0\}$; and the maximum itself is the expression $h_k(\bar{s}_k, \bar{a}_{k-1}) + C_k(\bar{s}_k, \bar{a}_{k-1}) I\{C_k(\bar{s}_k, \bar{a}_{k-1}) > 0\}$.

The premise of A -learning is thus to model the contrast functions rather than the full Q -functions as in Q -learning. For $k = K - 1, \dots, 1$ the latter involve possibly complex relationships, raising concern over the consequences of model misspecification for estimation of the optimal regime. As identifying the optimal regime depends only on correct specification of the contrast functions, A -learning may be less sensitive to misspecification.

We now describe the A -learning procedure. Assume posited models $C_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$, $k = 1, \dots, K$, say, for the contrast functions, each depending on a parameter ψ_k . Consider the K th decision. Given $C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K)$, letting $\pi_K(\bar{s}_K, \bar{a}_{K-1}) = \text{pr}(A_K = 1 | \bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1})$ be the propensity of receiving treatment 1 in the observed data as a function of past history and writing $\tilde{V}_{(K+1)i} = Y_i$, [Robins \(2004\)](#) showed that all consistent and asymptotically normal estimators for ψ_K are solutions to estimating equations of the form

$$\sum_{i=1}^n \lambda_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i}) \{A_{Ki} - \pi_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i})\} \times \{\tilde{V}_{(K+1)i} - A_{Ki} C_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i}; \psi_K) - \theta_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i})\} = 0 \quad (27)$$

for arbitrary functions $\lambda_K(\bar{s}_K, \bar{a}_{K-1})$ of the same dimension as ψ_K and $\theta_K(\bar{s}_K, \bar{a}_{K-1})$. Assuming the model $C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K)$ is correct, if $\text{var}(Y | \bar{S}_K = s_K, \bar{A}_{K-1} = a_{K-1})$ is constant, the optimal choices of these functions are $\lambda_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K) = \partial / \partial \psi_K C_K(\bar{s}_K, \bar{a}_{K-1}; \psi_K)$ and $\theta_K(\bar{s}_K, \bar{a}_{K-1}) = h_K(\bar{s}_K, \bar{a}_{K-1})$; otherwise, the optimal λ_K is complex ([Robins, 2004](#)).

To implement estimation of ψ_K via (27), one may adopt parametric models for these functions. Although the appeal of A -learning is that it obviates the need to specify fully the Q -functions, one may posit a model for the optimal θ_K , $h_K(\bar{s}_K, \bar{a}_{K-1}; \beta_K)$, say. Moreover, unless the data are from a SMART study, in which case the propensities $\pi_K(\bar{s}_K, \bar{a}_{K-1})$ would be known, these may also be modeled as $\pi_K(\bar{s}_K, \bar{a}_{K-1}; \phi_K)$ (e.g., by a logistic regression). These models are only adjuncts to estimating the parameter of interest, ψ_K ; interestingly, as long as at least one of these models is correctly specified, (27) will yield a consistent estimator for ψ_K , the so-called double robustness property. Substituting these models in (27), one solves

(27) jointly in $(\psi_K^T, \beta_K^T, \phi_K^T)^T$ with

$$\sum_{i=1}^n \frac{\partial h_K(\bar{S}_K, \bar{A}_{K-1}; \beta_K)}{\partial \beta_K} \{ \tilde{V}_{(K+1)i} - A_{Ki} C_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i}; \psi_K) - h_K(\bar{S}_{Ki}, \bar{A}_{(K-1)i}; \beta_K) \} = 0$$

and the usual binary regression likelihood score equations in ϕ_K . We then have $d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1}; \psi_K) = I[C_K\{\bar{s}_K, \bar{a}_{K-1}; \psi_K\} > 0]$; as in Q -learning, substituting $\hat{\psi}_K$ yields an estimator for the optimal treatment choice at decision K for a patient with past history $\bar{S}_K = s_K, \bar{A}_{K-1} = \bar{a}_{K-1}$.

With $\hat{\psi}_K$ in hand, as with Q -learning, the A -learning algorithm proceeds in a backward iterative fashion to yield $\hat{\psi}_k, k = K-1, \dots, 1$. At the k th decision, given models $h_k(\bar{s}_k, \bar{a}_{k-1}; \beta_k)$ and $\pi_k(\bar{s}_k, \bar{a}_{k-1}; \phi_k)$, one solves jointly in $(\psi_k^T, \beta_k^T, \phi_k^T)$ a system of estimating equations analogous to those above. As in Q -learning, the k th set of equations is based on “responses” $\tilde{V}_{(k+1)i}$, where, for each i , \tilde{V}_{ki} estimates $V_k(\bar{S}_{ki}, \bar{A}_{(k-1)i})$. It may be shown (see Section A.3 of the Appendix) that $E \left(V_{k+1}(\bar{S}_{k+1}, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1}) [I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k] \mid \bar{S}_k, \bar{A}_{k-1} \right) = V_k(\bar{S}_k, \bar{A}_{k-1})$. The expression $C_k(\bar{S}_k, \bar{A}_{k-1}) [I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k]$ is referred to as the advantage or regret function (Murphy, 2003), as it represents the “advantage” in response incurred if the optimal treatment at the k th decision were given relative to that actually received (or, equivalently, the “regret” incurred by not using the optimal treatment). Accordingly, define recursively $\tilde{V}_{ki} = \tilde{V}_{(k+1)i} + C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \hat{\psi}_k) [I\{C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \hat{\psi}_k) > 0\} - A_{ki}]$, $k = K, K-1, \dots, 1$, $\tilde{V}_{(K+1)i} = Y_i$. The equations at the k th decision are then

$$\begin{aligned} \sum_{i=1}^n \lambda_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \psi_k) \{ A_{ki} - \pi_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \phi_k) \} \\ \times \{ \tilde{V}_{(k+1)i} - A_{ki} C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \psi_k) - h_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \beta_k) \} = 0, \\ \sum_{i=1}^n \frac{\partial h_k(\bar{S}_k, \bar{A}_{k-1}; \beta_k)}{\partial \beta_k} \{ \tilde{V}_{(k+1)i} - A_{ki} C_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \psi_k) - h_k(\bar{S}_{ki}, \bar{A}_{(k-1)i}; \beta_k) \} = 0, \end{aligned} \quad (28)$$

for a given specification $\lambda_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$, solved jointly with the maximum likelihood score equations for binary regression in ϕ_k . It follows that $d_k^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}; \hat{\psi}_k) = I[C_k\{\bar{s}_k, \bar{a}_{k-1}; \hat{\psi}_k\} > 0]$. As above, the optimal λ_k is complex Robins (2004); taking $\lambda_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k) = \partial / \partial \psi_k C_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$ is reasonable for practical implementation.

Summarizing, the estimated optimal regime $\hat{d}_A^{\text{opt}} = (\hat{d}_{A,1}^{\text{opt}}, \dots, \hat{d}_{A,K}^{\text{opt}})$ is

$$\hat{d}_{A,1}^{\text{opt}}(s_1) = d_1^{\text{opt}}(s_1; \hat{\psi}_1), \quad \hat{d}_{A,k}^{\text{opt}}(\bar{s}_k, \bar{a}_{k-1}) = d_k^{\text{opt}}(\bar{s}_k, a_{k-1}; \hat{\psi}_k), \quad k = 2, \dots, K, \quad (29)$$

As with Q -learning, how well \hat{d}_A^{opt} estimates d^{opt} depends on how close the $C_k(\bar{s}_k, \bar{a}_{k-1}; \psi_k)$ are to the true contrast functions.

4.3 Comparison and Practical Considerations

When $K = 1$, the Q -function is a model for $E(Y|S_1 = s_1, A_1 = a_1)$. If in Q -learning this model and the variance model Σ_1 in (23) are correctly specified, then, as noted above, the form of (23) is optimal for estimating ξ_1 . Accordingly, even if $C_1(s_1; \psi_1)$ and $h_1(s_1; \beta_1)$ are correctly modeled, (28) with $K = 1$ is generally not of this optimal form for any choice $\lambda_1(s_1; \psi_1)$, and hence A -learning will yield relatively inefficient inference on ψ_1 and hence on the optimal regime. However, if in Q -learning the Q -function is misspecified, but in A -learning $C_1(s_1; \psi_1)$ and $\pi_1(s_1; \phi_1)$ are both correctly specified, then A -learning will still yield consistent inference on ψ_1 and hence the optimal regime, whereas inference on ξ_1 and the optimal regime via Q -learning may be inconsistent. We assess the trade-off between consistency and efficiency in this case in Section 5. For $K > 1$, owing to the complications involved in specifying optimal estimating equations for Q - and A -learning, the relative performance of the methods is not readily apparent; we investigate in Section 5.

In certain special cases, Q - and A -learning lead to identical estimators for the Q -function (Chakraborty et al., 2010). For example, this holds if the propensities for treatment are constant, as would be the case under pure randomization at each decision point, and certain linear models are used for $C_1(s_1; \psi_1)$ and $h_1(s_1; \beta_1)$; see Section A.4 of the Appendix for a demonstration when $K = 1$ and $\text{pr}(A_1 = 1|S_1 = s_1)$ does not depend on s_1 .

As we have emphasized, for Q -learning, while modeling the Q -function at decision K is a standard regression problem with response Y , for decisions $K - 1, \dots, 1$, this involves modeling the estimated value function, which depends on the relationships for subsequent decisions. Ideally, the sequence of posited models $Q_k(\bar{s}_k, \bar{a}_k; \xi_k)$ should respect this constraint. However, this may be difficult to achieve with standard

regression models. To illustrate, consider the models in (26), and assume S_1, S_2 are scalar, where the conditional distribution of S_2 given $S_1 = s_1, A_1 = a_1$ is $\text{Normal}(\mathcal{K}_1^T \gamma, \sigma^2)$, say, $\mathcal{K}_1 = (1, s_1, a_1)^T$. Recall that $V_2(\bar{s}_2, a_1; \xi_2) = \mathcal{H}_2^T \beta_2 + (\mathcal{H}_2^T \psi_2) I(\mathcal{H}_2^T \psi_2 > 0)$, where we can write $\mathcal{H}_2^T \beta_2 = \mathcal{K}_1^T \beta_{21} + s_2 \beta_{22}$ and $\mathcal{H}_2^T \psi_2 = \mathcal{K}_1^T \psi_{21} + s_2 \psi_{22}$. Then, if the model Q_2 in (26) were correct, from (17), ideally, $Q_1(s_1, a_1) = \mathbb{E}\{V_2(s_1, S_2, a_1; \xi_2) | S_1 = s_1, A_1 = a_1\}$. Letting $\varphi(\cdot)$ and $\Phi(\cdot)$ be the standard normal density and cumulative distribution function, respectively, it may be shown (see Section A.5 of the Appendix) that, under these conditions,

$$\begin{aligned} Q_1(s_1, a_1) &= \mathbb{E}\{V_2(s_1, S_2, a_1; \xi_2) | S_1 = s_1, A_1 = a_1\} = \mathcal{K}_1^T (\beta_{21} + \gamma \beta_{22}) \\ &\quad + (\mathcal{K}_1^T \psi_{21}) \{1 - \Phi(\eta)\} + \psi_{22} \{\sigma \varphi(\eta) + (\mathcal{K}_1^T \gamma) \{1 - \Phi(\eta)\}\}, \end{aligned} \quad (30)$$

where $\eta = -\mathcal{K}_1^T (\psi_{21}/\psi_{22} + \gamma)/\sigma$, and we have taken $\psi_{22} > 0$. Contrast the implied true $Q_1(s_1, a_1)$ in (30) to the posited linear model in (26); clearly, the true relationship is highly nonlinear in s_1, a_1 and is likely to be poorly approximated by $Q_1(s_1, a_1; \xi_1)$ in (26). Evidently, for larger K , this incompatibility between true and assumed models would propagate from $K - 1, \dots, 1$. Thus, while the use of linear models for the Q -functions is popular in practice, the potential for such mismodeling should be recognized.

An alternative approach that may mitigate the risk of mismodeling is to employ flexible models for the Q -functions. Zhao, Kosorok, and Zeng (2009) use support vector regression models in place of the linear models described above. Indeed, recent developments in statistical learning suggest a large collection of powerful regression methods that might be used. Many of these methods must be tuned in order to balance bias and variance, a natural approach to which is to minimize the cross-validated mean squared error of the Q -functions at each decision point. An obvious downside is that the final model may be difficult to interpret, and clinicians may be unwilling to implement “black box” rules. One compromise is to fit a simple, interpretable model, such as a decision tree, to the fitted values of the complex model in order to get a feel for what factors are driving the recommended treatment decisions. One can then check the simple model against scientific theory. If the simple, approximate model appears sensible, then clinicians may be willing to use predictions from the more complex and less interpretable model. For further discussion and references, see Craven and Shavlik (1996).

A -learning represents a middle ground between Q -learning and these approaches in that it allows for flexible modeling of the functions $h_k(\bar{s}_k, \bar{a}_{k-1})$ while maintaining simple parametric models for the contrast functions $C_k(\bar{s}_k, \bar{a}_{k-1})$. Thus, the resulting decision rule, which depends only on the contrast function, remains interpretable, while the model for the response is allowed to be nonlinear. This is also appealing in that it may be reasonable to expect, based on the underlying science, that the relationship between patient history and outcome is complex while the optimal rule for treatment assignment is dependent, in a simple fashion, on a small number of variables. The flexibility allowed by a semi-parametric model also has its drawbacks. Techniques for formal model building, critique, and diagnosis are well understood for linear models but much less so for semi-parametric models. Consequently, Q -learning based on building a series of linear models may be more appealing to an analyst interested in formal diagnostics.

5 Simulation Studies

We examine the finite sample performance of Q - and A -learning on a suite of test examples via Monte Carlo simulation. To illustrate the trade-offs between the methods discussed in the preceding sections, we begin with correctly specified models and then systematically introduce misspecification of the Q -function, the propensity model, and both the Q -function and propensity model. In all cases, the contrast function is correctly specified, as, when this is not the case, the form of the optimal regime induced by an incorrect contrast function may not include d^{opt} , making interpretation difficult. In all scenarios, 10,000 Monte Carlo replications were used, and, for each generated data set, the estimated optimal regimes \hat{d}_Q^{opt} and \hat{d}_A^{opt} in (25) and (29) were obtained using the Q - and A -learning procedures described in Sections 4.1 and 4.2.

For simplicity, we consider one ($K = 1$) and two stage ($K = 2$) decision problems, where, at each decision point, there are two feasible treatment options coded as 0 and 1. In all cases, we used Q -functions of the form $Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1 C_1(s_1; \psi_1)$ and $Q_2(\bar{s}_2, \bar{a}_2; \xi_2) = h_2(\bar{s}_2; a_1; \beta_2) + a_2 C_2(\bar{s}_2, a_1; \psi_2)$ to represent both true and assumed working models. With the contrast functions correctly specified, the parameters ψ_k , $k = 1, 2$, dictate the optimal regime. Thus, as one measure of performance, we focus on relative efficiency of the estimators of components of ψ_k obtained by Q -learning to those obtained by A -learning, as reflected by the ratio of their Monte Carlo mean squared errors (MSEs) (so by MSE of A -

learning/MSE of Q -learning), so that values greater than 1 favor Q -learning. Recognizing that $E\{Y^*(d^{\text{opt}})\}$ is the benchmark achievable outcome on average, as a second measure, we consider the extent to which the estimated regimes \hat{d}_Q^{opt} and \hat{d}_A^{opt} achieve $E\{Y^*(d^{\text{opt}})\}$ if followed by the population. Specifically, for regime d indexed by ψ_1 ($K = 1$) or ψ_1 and ψ_2 ($K = 2$), let $H(d) = E\{Y^*(d)\}$, a function of these parameters. Then $H(d^{\text{opt}}) = E\{Y^*(d^{\text{opt}})\}$ is this function evaluated at the true parameter values, and $H(\hat{d}^{\text{opt}})$ is this function evaluated the estimated parameter values for a given data set, where \hat{d}^{opt} represents \hat{d}_Q^{opt} or \hat{d}_A^{opt} . Define $R(\hat{d}^{\text{opt}}) = E\{H(\hat{d}^{\text{opt}})\}/H(d^{\text{opt}})$, where the expectation in the numerator is with respect to the distribution of the estimated parameters in \hat{d}^{opt} , which may be interpreted as reflecting the efficiency with which \hat{d}^{opt} achieves the performance of the true optimal regime. In Section A.6 of the Appendix, we discuss calculation of $R(\hat{d}^{\text{opt}})$.

5.1 One Decision Point

In this and the next section, $n = 200$. Here, the observed data are (S_{1i}, A_{1i}, Y_i) , $i = 1, \dots, n$. With $\text{expit}(x) = e^x/(1 + e^x)$, to generate the data, we used

$$\begin{aligned} S_1 &\sim \text{Normal}(0, 1), \quad A_1|S_1 = s_1 \sim \text{Bernoulli}\{\text{expit}(\phi_{10}^0 + \phi_{11}^0 s_1 + \phi_{12}^0 s_1^2)\}, \\ Y|S_1 = s_1, A_1 = a_1 &\sim \text{Normal}\{\beta_{10}^0 + \beta_{11}^0 s_1 + \beta_{12}^0 s_1^2 + a_1(\psi_{10}^0 + \psi_{11}^0 s_1), 3\}, \end{aligned}$$

so that the class of generative models is indexed by $\theta^0 = (\phi_{10}^0, \phi_{11}^0, \phi_{12}^0, \beta_{10}^0, \beta_{11}^0, \beta_{12}^0, \psi_{10}^0, \psi_{11}^0)^T$, and $d^{\text{opt}} = d_1^{\text{opt}}$, $d_1^{\text{opt}}(s_1) = I(\psi_{10}^0 + \psi_{11}^0 s_1 > 0)$. For A -learning, we assumed working models $h_1(s_1; \beta_1) = \beta_{10} + \beta_{11} s_1$, $C_1(s_1; \psi_1) = \psi_{10} + \psi_{11} s_1$, and $\pi_1(s_1; \phi_1) = \text{expit}(\phi_{10} + \phi_{11} s_1)$, and for Q -learning used $Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1 C_1(s_1; \psi_1)$. Note that these working models involve correctly specified contrast functions and are nested within the true generative models, with $h_1(s_1; \beta_1)$, and hence the Q -function, correctly specified when $\beta_{12}^0 = 0$. Similarly, the propensity model $\pi_1(s_1; \phi_1)$ is correctly specified when $\phi_{12}^0 = 0$. To study the effects of misspecification, we systematically varied these two parameters while keeping the others fixed, considering parameter settings of the form $\theta^0 = (0, -2, \phi_{12}^0, 1, 1, \beta_{12}^0, 1, 0.5)^T$.

Correctly specified models. As noted in Section 4.3, when all working models are correctly specified, Q -learning is more efficient than A -learning. Under our class of generative models, this occurs when

$\beta_{12}^0 = \phi_{12}^0 = 0$. In this scenario, the relative efficiency of Q -learning with respect to A -learning is 1.06 for estimating ψ_{10}^0 and 2.74 for estimating ψ_{11}^0 . Thus, Q -learning is a modest 6% more efficient in estimating ψ_{10}^0 but a dramatic 174% more efficient in estimating ψ_{11}^0 . Interestingly, the efficiency of the decision rules produced by Q - and A -learning is similar, with $R(\hat{d}_Q^{\text{opt}}) = 0.97$ and $R(\hat{d}_A^{\text{opt}}) = 0.95$, so that the relative inefficiency in estimation of ψ_1 suffered by A -learning does not translate in to a regime of poorer quality than that found via Q -learning.

Misspecified propensity model. An appeal of A -learning is the double robustness property noted in Section 4.2, which implies that ψ_1 should be estimated consistently when the propensity model is misspecified provided that the Q -function is correct. Under our class of generative models, this corresponds to $\beta_{12}^0 = 0$ and nonzero ϕ_{12}^0 . In contrast, Q -learning does not depend on the propensity model, so its performance is unaffected by this misspecification. Figure 1 shows the relative efficiency in estimating ψ_{10}^0 and ψ_{11}^0 and the efficiency of \hat{d}_Q^{opt} and \hat{d}_A^{opt} as ϕ_{12}^0 varies from -1 to 1 . The leftmost panel shows that there is minimal gain in efficiency by using Q -learning instead of A -learning in estimation of ψ_{10}^0 . From the center panel, Q -learning yields substantial gains over A -learning for estimating ψ_{11}^0 . Interestingly, the gain in efficiency of Q - over A -learning is largest when $\phi_{12}^0 = 0$, which corresponds to the propensity model being correctly specified. Letting $\pi^0(s_1; \phi_1^0)$ be the true propensity, $\phi_1^0 = (\phi_{10}^0, \phi_{11}^0, \phi_{12}^0)^T$, a possible explanation for this seemingly contradictory result is that, as $|\phi_{12}^0|$ gets larger, $\text{logit}\{\pi^0(S_1; \phi_1^0)\} = \phi_{10}^0 + \phi_{11}^0 s_1 + \phi_{12}^0 s_1^2$ becomes more profoundly quadratic. Consequently, the estimator for ϕ_{11} in the posited model $\pi_1(s_1; \phi_1) = \text{expit}(\phi_{10} + \phi_{11} s_1)$ approaches zero, so that the posited propensity approaches a constant. Because Q - and A -learning are equivalent under constant propensity, the efficiency gains decrease as $|\phi_{12}^0| \rightarrow \infty$. The right panel of Figure 1 shows a small gain in efficiency of \hat{d}_Q^{opt} over \hat{d}_A^{opt} , with both achieving good performance.

Misspecified Q -function. This scenario examines the second aspect of A -learning’s double-robustness and is characterized in our class of true generative models by $\phi_{12}^0 = 0$ and nonzero β_{12}^0 . Here, A -learning leads to consistent estimation while Q -learning need not. The left panel of Figure 2 shows that the gain in efficiency using A -learning is minimal in estimating ψ_{10}^0 . The center panel illustrates the bias-variance trade-off associated with choice between Q - and A -learning. For values of β_{12}^0 that are far from zero, the bias in the misspecified Q -function dominates the variance, and A -learning enjoys smaller MSE while, for

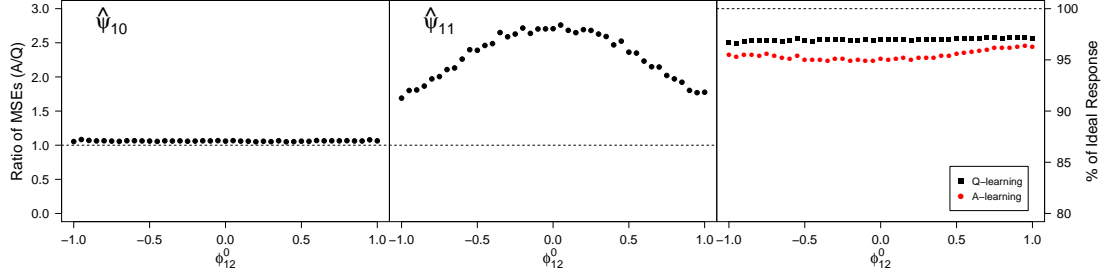


Figure 1: Monte Carlo MSE ratios for estimators of components of ψ_1 (left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (right panel) under misspecification of the propensity model. MSE ratios > 1 favor Q -learning

small values of β_{12}^0 , variance dominates bias, and Q -learning is more efficient. The right panel shows that large bias in the Q -function can lead to meaningful loss (around 10%) in efficiency of \hat{d}_Q^{opt} relative to \hat{d}_A^{opt} .

Both propensity model and Q -function misspecified. In our class of generative models, this corresponds to nonzero values of both β_{12}^0 and ϕ_{12}^0 . Rather than vary both values, (e.g., over a grid), we varied one and chose the other so that it is “equivalently misspecified.” In particular, for a given value of ϕ_{12}^0 , we selected $\beta_{12}^0 = \beta_{12}^0(\phi_{12}^0)$ so that the t -statistic associated with testing $\phi_{12}^0 = 0$ in the logistic propensity model and the t -statistic associated with testing $\beta_{12}^0 = 0$ in the linear Q -function would be approximately equal in distribution. Consequently, across data sets, an analyst would be equally likely to detect either form of misspecification. Details of this construction are given in Section A.7 of the Appendix.

As in the preceding scenario, Figure 3 illustrates the bias-variance trade-off associated with Q - and A -learning. For large misspecification, A -learning provides a large enough reduction in bias to yield lower MSE; for small misspecification, Q -learning incurs some bias but reduces the variance enough to yield lower MSE. From the right panel of the figure, bias seems to translate into a larger loss in quality of the estimators of d^{opt} than variance.

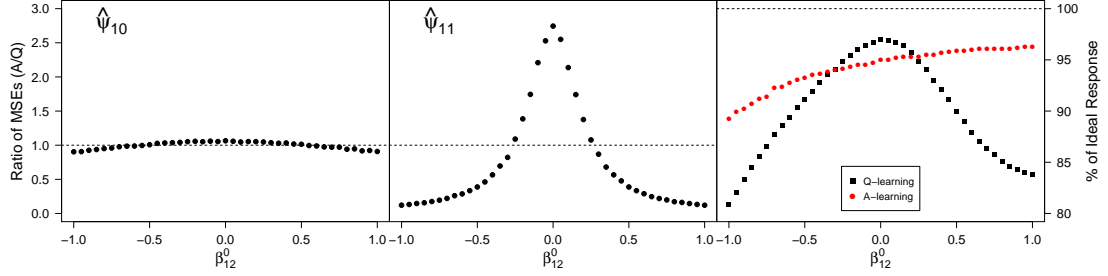


Figure 2: Monte Carlo MSE ratios for estimators of components of ψ_1 (left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (right panel) under misspecification of the Q -function. MSE ratios > 1 favor Q -learning

5.2 Two Decision Points

For $K = 2$, the observed data available to estimate $d^{\text{opt}} = (d_1^{\text{opt}}, d_2^{\text{opt}})$ are $(S_{1i}, A_{1i}, S_{2i}, A_{2i}, Y_i), i = 1, \dots, n$. For these scenarios, we used a class of true generative data models that differs from those of [Chakraborty et al. \(2010\)](#), [Song et al. \(2010\)](#), and [Laber et al. \(2010\)](#) only in that S_2 is continuous instead of binary. The generative model is

$$S_1 \sim \text{Bernoulli}(0.5), \quad A_1|S_1 = s_1 \sim \text{Bernoulli}\{\text{expit}(\phi_{10}^0 + \phi_{11}^0 s_1)\},$$

$$S_2|S_1 = s_1, A_1 = a_1 \sim \text{Normal}(\delta_{10}^0 + \delta_{11}^0 s_1 + \delta_{12}^0 a_1 + \delta_{13}^0 s_1 a_1, 2),$$

$$A_2|S_1 = s_1, S_2 = s_2, A_1 = a_1 \sim \text{Bernoulli}\{\text{expit}(\phi_{20}^0 + \phi_{21}^0 s_1 + \phi_{22}^0 a_1 + \phi_{23}^0 s_2 + \phi_{24}^0 a_1 s_2 + \phi_{25}^0 s_2^2)\},$$

$$Y|S_1 = s_1, S_2 = s_2, A_1 = a_1, A_2 = a_2 \sim \text{Normal}\{m(s_1, s_2, a_1, a_2), 10\},$$

$$m(s_1, s_2, a_1, a_2) = \beta_{20}^0 + \beta_{21}^0 s_1 + \beta_{22}^0 a_1 + \beta_{23}^0 s_1 a_1 + \beta_{24}^0 s_2 + \beta_{25}^0 s_2^2 + a_2(\psi_{20}^0 + \psi_{21}^0 a_1 + \psi_{22}^0 s_2).$$

The model is indexed by $\phi_1^0 = (\phi_{10}^0, \phi_{11}^0)^T$, $\delta_1^0 = (\delta_{10}^0, \delta_{11}^0, \delta_{12}^0, \delta_{13}^0)^T$, $\phi_2^0 = (\phi_{20}^0, \phi_{21}^0, \phi_{22}^0, \phi_{23}^0, \phi_{24}^0, \phi_{25}^0)^T$, $\beta_2^0 = (\beta_{20}^0, \beta_{21}^0, \beta_{22}^0, \beta_{23}^0, \beta_{24}^0, \beta_{25}^0)^T$, and $\psi_2^0 = (\psi_{20}^0, \psi_{21}^0, \psi_{22}^0)^T$, with true $h_2^0(s_1, s_2, a_1) = \beta_{20}^0 + \beta_{21}^0 s_1 + \beta_{22}^0 a_1 + \beta_{23}^0 s_1 a_1 + \beta_{24}^0 s_2 + \beta_{25}^0 s_2^2$ and contrast function $C_2^0(s_1, s_2, a_1) = \psi_{20}^0 + \psi_{21}^0 a_1 + \psi_{22}^0 s_2$, say. Because

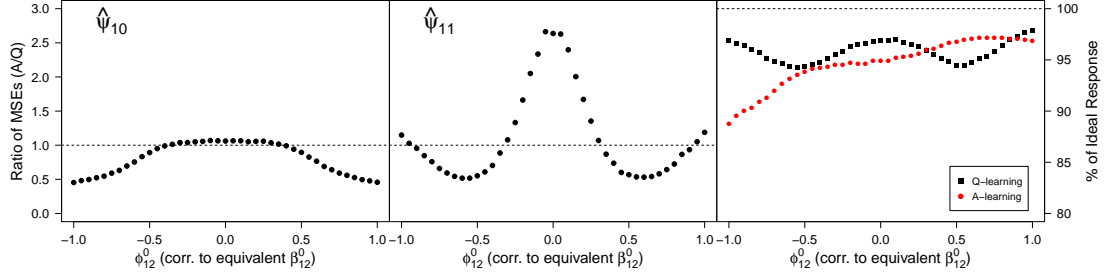


Figure 3: Monte Carlo MSE ratios for estimators of components of ψ_1 (left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (right panel) under misspecification of both the propensity model and the Q -function. MSE ratios > 1 favor Q -learning

A_1 and S_1 are binary, the true functions $h_1^0(s_1) = \beta_{10}^0 + \beta_{11}^0 s_1$ and $C_1^0(s_1) = \psi_{10}^0 + \psi_{11}^0 s_1$, are linear in s_1 ; $\beta_{10}^0, \beta_{11}^0, \psi_{10}^0$, and ψ_{11}^0 are derived in terms of parameters indexing the generative model in Section A.8 of the Appendix. Thus, the true optimal regime has $d_1^{\text{opt}}(s_1) = I(\psi_{10}^0 + \psi_{11}^0 s_1 > 0)$ and $d_2^{\text{opt}}(s_1, s_2, a_1) = I(\psi_{20}^0 + \psi_{21}^0 a_1 + \psi_{22}^0 s_2 > 0)$.

We assumed working models for A -learning of the form $h_1(s_1; \beta_1) = \beta_{10} + \beta_{11} s_1$, $C_1(s_1; \psi_1) = \psi_{10} + \psi_{11} s_1$, $\pi_1(s_1; \phi_1) = \text{expit}(\phi_{10} + \phi_{11} s_1)$, $h_2(s_1, s_2, a_1; \beta_2) = \beta_{20} + \beta_{21} s_1 + \beta_{22} a_1 + \beta_{23} s_1 a_1 + \beta_{24} s_2$, $C_2(s_1, s_2, a_1; \psi_2) = \psi_{20} + \psi_{21} a_1 + \psi_{22} s_2$, and $\pi_2(s_1, s_2, a_1; \phi_2) = \text{expit}(\phi_{20} + \phi_{21} s_1 + \phi_{22} a_1 + \phi_{23} s_2 + \phi_{24} a_1 s_2)$; and, similarly, assumed Q -functions of the form $Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1 C_1(s_1; \psi_1)$ and $Q_2(s_1, s_2, a_1, a_2; \xi_2) = h_2(s_1, s_2, a_1; \beta_2) + a_2 C_2(s_1, s_2, a_1; \psi_2)$ for Q -learning, so that the contrast functions are correctly specified in each case. Comparison of the working and generative models shows that the former are correctly specified when ϕ_{25}^0 and β_{25}^0 are both zero and are misspecified otherwise. Thus, we systematically varied these parameters to study the effects of misspecification, leaving all other parameter values fixed, taking $\phi_1^0 = (0.3, -0.5)^T$, $\delta_1^0 = (0, 0.5, -0.75, 0.25)^T$, $\phi_2^0 = (0, 0.5, 0.1, -1, -0.1, \phi_{25}^0)^T$, $\beta_2^0 = (3, 0, 0.1, -0.5, -0.5, \beta_{25}^0)^T$, and $\psi_2^0 = (1, 0.25, 0.5)^T$.

Correctly specified models. Given our working models, this occurs when $\phi_{25}^0 = \beta_{25}^0 = 0$ in the generative models. As discussed previously, Q -learning is efficient when the models are correctly specified. Relative efficiencies of Q -learning with respect to A -learning for estimating $\psi_{10}^0, \psi_{11}^0, \psi_{20}^0, \psi_{21}^0$, and ψ_{22}^0 are 1.07, 1.03,

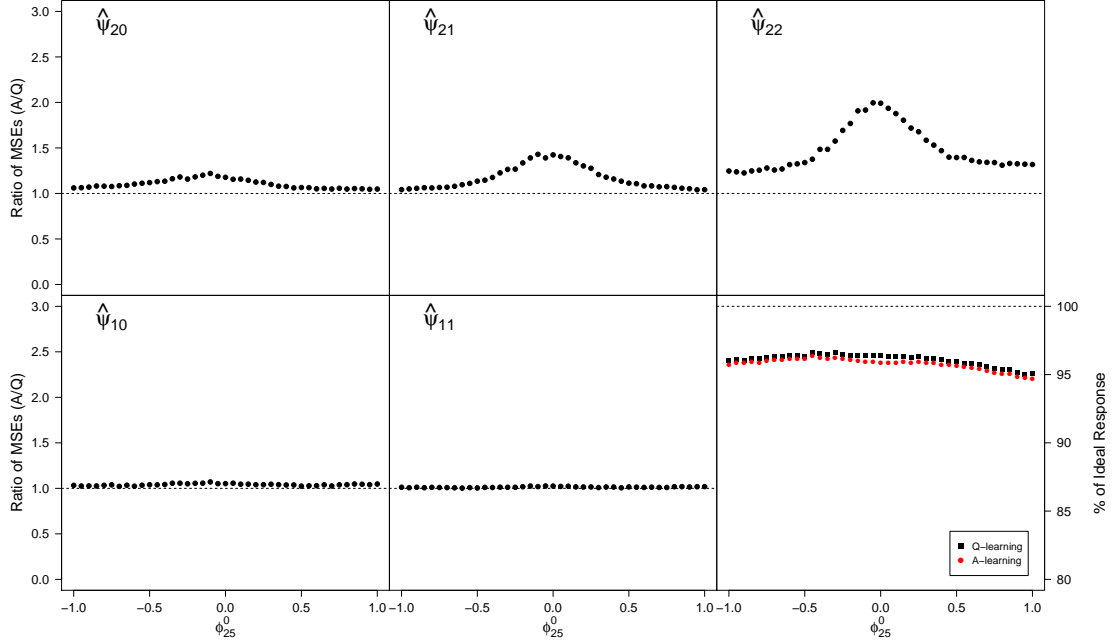


Figure 4: Monte Carlo MSE ratios for estimators of components of ψ_2 and ψ_1 (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (lower right panel) under misspecification of the propensity model. MSE ratios > 1 favor Q -learning

1.19, 1.44, and 1.98, respectively. Hence, Q -learning is markedly more efficient in estimating the second stage parameters but only modestly so in estimating first stage parameters. More efficient estimators of the underlying parameters do not translate into significantly more efficient estimated regimes, as $R(\hat{d}_Q^{\text{opt}}) = 0.96$ and $R(\hat{d}_A^{\text{opt}}) = 0.96$.

Misspecified propensity model. The propensity model at the second stage is misspecified when ϕ_{25}^0 is nonzero. To isolate the effects of such misspecification, we set $\beta_{25}^0 = 0$ and varied ϕ_{25}^0 between -1 and 1 . From Figure 4, Q -learning is more efficient than A -learning for estimation of all parameters in ψ_1 and ψ_2 , and, as in the one decision case, the efficiency gain is largest when the $\phi_{25}^0 = 0$, corresponding to a correctly specified propensity model. From the lower right panel, there appears to be little difference in efficiency of \hat{d}_Q^{opt} and \hat{d}_A^{opt} .

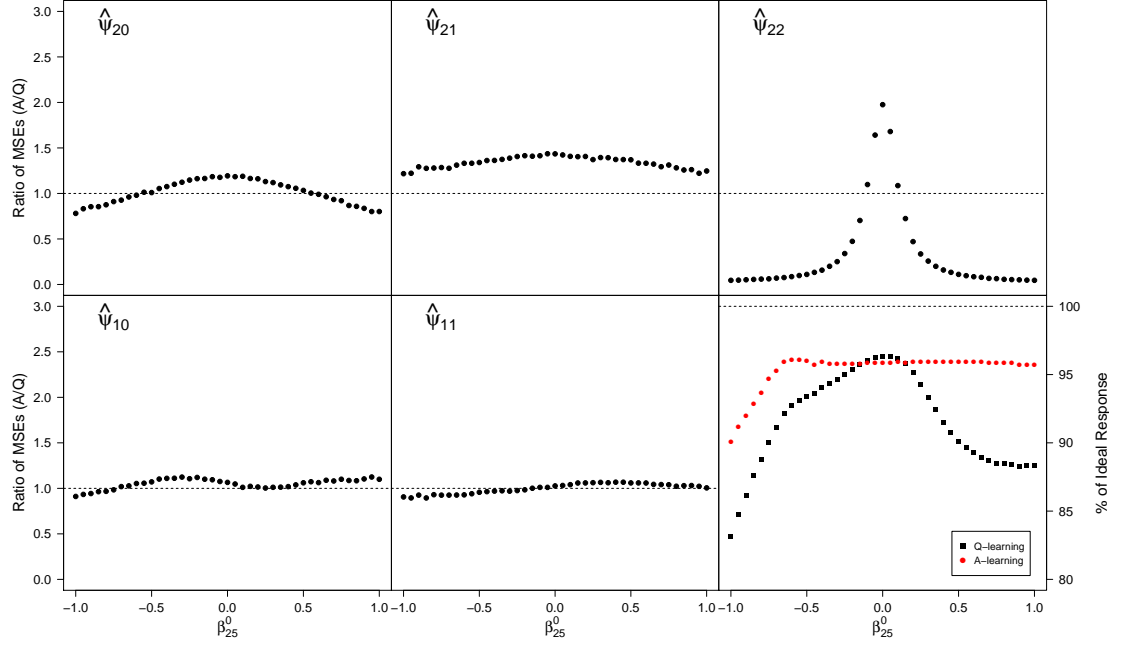


Figure 5: Monte Carlo MSE ratios for estimators of components of ψ_2 and ψ_1 (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (lower right panel) under misspecification of the Q -functions. MSE ratios > 1 favor Q -learning

Misspecified Q -function. Under our class of generative models, the Q -function is misspecified when β_{25}^0 is nonzero. We set $\phi_{25}^0 = 0$ to focus on the effects of such misspecification. Figure 5 shows that, for the first stage parameters ψ_{10}^0 and ψ_{11}^0 , there is little difference in efficiency between Q - and A -learning. The upper panels illustrate varying degrees of the bias-variance trade-off between the methods. In particular, in estimating ψ_{22}^0 , a small amount of misspecification leads to significant bias, and hence A -learning produces a much more accurate estimator, while, for ψ_{20}^0 , the bias-variance trade-off is present but attenuated, and there is little difference between Q - and A -learning. In estimation of ψ_{21}^0 , variance appears to dominate bias, and Q -learning is preferred for the chosen range of β_{25}^0 values. From the lower right panel, relative efficiency for estimating ψ_{22}^0 weakly tracks the relative efficiencies of the estimated regimes \hat{d}_Q^{opt} and \hat{d}_A^{opt} , suggesting that the efficiency gain for A -learning in estimating ψ_{22}^0 leads to improved estimation of d^{opt} .

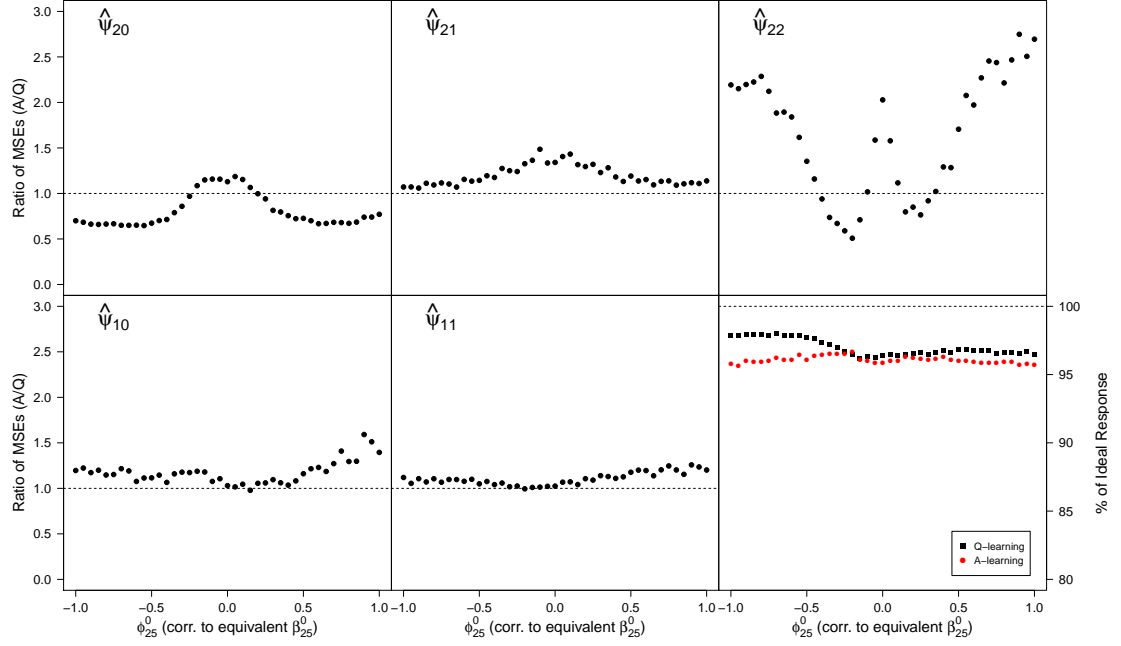


Figure 6: Monte Carlo MSE ratios for estimators of components of ψ_2 and ψ_1 (upper row and lower row left and center panels) and efficiencies $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ for estimating the true d^{opt} (lower right panel) under misspecification of both the propensity models and Q -functions. MSE ratios > 1 favor Q -learning

Both the propensity model and Q -function misspecified. Under our generative model, this scenario corresponds to nonzero values of β_{25}^0 and ϕ_{25}^0 . Analogous to the one decision case, we chose pairs $(\beta_{25}^0, \phi_{25}^0)$ that are “equivalently misspecified;” see Section A.7 of the Appendix. Figure 6 shows the relative efficiency of the two methods. There is no general trend in efficiency of estimation across parameters that might recommend one method over the other. Furthermore, from the lower right panel, there is little difference in efficiency of the estimated regimes. This is as expected, as one should not expect to draw broad conclusions, as neither Q - nor A -learning need be consistent here. Interestingly, despite misspecification of both models, \hat{d}_Q^{opt} and \hat{d}_A^{opt} still enjoy high efficiency.

5.3 Moodie and Richardson Scenario

The foregoing simulation scenarios deliberately involve simple models for the Q -functions in order to allow straightforward interpretation. To investigate the relative performance of the methods in a more challenging setting, we generated data from a scenario similar to that in [Moodie et al. \(2007\)](#) in which the true contrast functions are simple yet the Q -functions are complex.

The data generating process used mimics a study in which HIV-infected patients are randomized to receive antiretroviral therapy (coded as 1) or not (coded as 0) at baseline and again at six months, where the randomization probabilities depend on baseline and six month CD4 counts. Specifically, we generated baseline CD4 count $S_1 \sim \text{Normal}(450, 150^2)$, and baseline treatment A_1 was then assigned according to $A_1|S_1 = s_1 \sim \text{Bernoulli}\{\text{expit}(\phi_{10}^0 + \phi_{11}^0 s_1)\}$. We generated six month CD4 count S_2 , distributed conditional on $S_1 = s_1, A_1 = a_1$ as $\text{Normal}(1.25s_1, 60^2)$. Treatment A_2 was then generated according to $A_2|S_1 = s_1, A_1 = a_1, S_2 = s_2 \sim \text{Bernoulli}\{\text{expit}(\phi_{20}^0 + \phi_{21}^0 s_2)\}$. In contrast to the scenario in [Moodie et al. \(2007\)](#), this allows all possible treatment combinations. The outcome Y is CD4 count at one year; following [Moodie et al. \(2007\)](#), Y was generated as $Y = Y^{\text{opt}} - \mu_1^0(S_1, A_1) - \mu_2^0(S_1, S_2, A_1, A_2)$, where $Y^{\text{opt}}|S_1 = s_1, A_1 = a_1, S_2 = s_2, A_2 = a_2 \sim \text{Normal}(400 + 1.6s_1, 60^2)$. Here, $\mu_1^0(S_1, A_1)$ and $\mu_2^0(S_1, S_2, A_1, A_2)$ are the true advantage (regret) functions; we took $C_1^0(s_1) = \psi_{10}^0 + \psi_{11}^0 s_1$ and $C_2^0(s_1, s_2, a_1) = \psi_{20}^0 + \psi_{21}^0 s_2$ to be the true contrast functions, so that, from [Section 4.2](#),

$$\mu_1^0(S_1, A_1) = (\psi_{10}^0 + \psi_{11}^0 S_1)\{I(\psi_{10}^0 + \psi_{11}^0 S_1 > 0) - A_1\}, \quad (31)$$

$$\mu_2^0(S_1, S_2, A_1, A_2) = (\psi_{20}^0 + \psi_{21}^0 S_2)\{I(\psi_{20}^0 + \psi_{21}^0 S_2 > 0) - A_2\}. \quad (32)$$

It follows that the optimal treatment regime $d^{\text{opt}} = (d_1^{\text{opt}}, d_2^{\text{opt}})$ has $d_1^{\text{opt}}(s_1) = I(\psi_{10}^0 + \psi_{11}^0 s_1 > 0)$ and $d_2^{\text{opt}}(s_1, s_2, a_1) = I(\psi_{20}^0 + \psi_{21}^0 s_2 > 0)$. While the true contrast functions are linear in ψ_k^0 , $k = 1, 2$, the true implied $h_1^0(s_1)$ and $h_2^0(s_1, a_1, s_2)$ are nonsmooth and possibly complex.

Following [Moodie et al. \(2007\)](#), for A -learning, we assumed working models $h_1(s_1; \beta_1) = \beta_{10} + \beta_{11}s_1$, $C_1(s_1; \psi_1) = \psi_{10} + \psi_{11}s_1$, $h_2(s_1, s_2, a_1; \beta_2) = \beta_{20} + \beta_{21}s_1 + \beta_{22}a_1 + \beta_{23}s_1a_1 + \beta_{24}s_2$, and $C_2(s_1, s_2, a_1; \psi_2) = \psi_{20} + \psi_{21}s_2$, and assumed propensity models of the form $\pi_1(s_1; \phi_1) = \phi_{10} + \phi_{11}s_1$ and $\pi_2(s_1, s_2, a_1; \phi_2) =$

$\phi_{20} + \phi_{21}s_2$. For Q -learning, we analogously assumed Q -functions $Q_1(s_1, a_1; \xi_1) = h_1(s_1; \beta_1) + a_1 C_1(s_1; \psi_1)$ and $Q_2(s_1, s_2, a_1, a_2; \xi_2) = h_2(s_1, s_1, a_1; \beta_2) + a_2 C_2(s_1, s_2, a_1; \psi_2)$. Note that the contrast functions in each case are correctly specified, as are the propensity models; however, the Q -functions are misspecified, as the linear models $h_1(s_1; \beta_1)$ and $h_2(s_1, s_1, a_1; \beta_2)$ are poor approximations to the complex forms of the true $h_1^0(s_1)$ and $h_2^0(s_1, s_2, a_1)$.

We report results for $n = 1000$ with $\phi_1^0 = (\phi_{10}^0, \phi_{11}^0)^T = (2.0, -0.006)^T$, $\phi_2^0 = (\phi_{20}^0, \phi_{21}^0)^T = (0.8, -0.004)^T$, $\psi_1^0 = (\psi_{10}^0, \psi_{11}^0)^T = (250, -1.0)^T$, and $\psi_2^0 = (\psi_{20}^0, \psi_{21}^0)^T = (720, -2.0)^T$ in Table 1. Because the Q -functions are misspecified, not unexpectedly, the Q -learning estimators for ψ_1^0 and ψ_2^0 are biased, while those obtained via A -learning are consistent owing to the double robustness property. This leads to the dramatic relative inefficiency of Q -learning reflected by the MSE ratios. Under the assumed models, the estimated optimal regime for Q -learning dictates that, at baseline, antiretroviral therapy be given to patients with baseline CD4 count less than 199.7, while that estimated using A -learning gives treatment to those with baseline CD4 count less than 249.1, almost perfectly achieving the true optimal CD4 threshold of 250. Under the data generative process, using the baseline decision rule estimated via Q -learning may result in as many as 4.4% of patients who would receive therapy at baseline under the true optimal regime being assigned no treatment. Similarly, at the second decision, the estimated optimal regimes obtained by Q - and A -learning dictate that therapy be given to patients with six month CD4 count less than 320.2 and 360.1, respectively. Again, A -learning yields an estimated threshold almost identical to the optimal value of 360. Although that obtained via Q -learning is lower, 4.3% of patients who should receive therapy at six months would not if the estimated six month rule from Q -learning were followed by the population.

Using the approach outlined in Section A.6 of the Appendix, we have $H(d^{\text{opt}}) = 1120$, whereas $E\{H(\hat{d}_Q^{\text{opt}})\} \approx 1117.1$ (estimated standard error 1.3) and $E\{H(\hat{d}_A^{\text{opt}})\} \approx 1119.9$ (0.3), so that $R(\hat{d}_Q^{\text{opt}})$ and $R(\hat{d}_A^{\text{opt}})$ are virtually equal to one. Thus, although Q -learning results in poor estimation of parameters in the contrast functions, efficiency loss for estimating the optimal regime is negligible. A possible explanation is that for the advantage (regret) functions in (31) and (32), patients near the true treatment decision boundary would have $C_k^0(\bar{S}_k, \bar{A}_{k-1})$, $k = 1, 2$, close to zero. Thus, even if a regime improperly assigns treatment, patients near this boundary have only a small loss in expected outcome. This, and

Table 1: Monte Carlo average (standard deviation) of estimates obtained via Q - and A -learning and ratio of Monte Carlo MSE for the Moodie and Richardson scenario; MSE ratios > 1 favor Q -learning

Parameter (true value)	Q -learning	A -learning	MSE ratio
$\psi_{10}^0 = 250$	154.8 (23.2)	249.1 (18.7)	0.036
$\psi_{11}^0 = -1.0$	-0.775 (0.052)	-0.998 (0.041)	0.032
$\psi_{20}^0 = 720$	507.3 (49.2)	720.3 (48.4)	0.050
$\psi_{21}^0 = -2.0$	-1.584 (0.092)	-2.001 (0.085)	0.040

the aforementioned fact that only a small subset of the population is affected by poor treatment decisions under Q -learning, results in the relatively good expected outcome under the estimated Q -learning regime.

6 Application to STAR*D

Sequenced Treatment Alternatives to Relieve Depression (STAR*D) was a prospective multisite, randomized clinical trial enrolling 4041 patients designed to compare various treatment options for patients with major depressive disorder. The trial involved four levels, where each level consisted of a 12 week period of treatment, with scheduled clinic visits at approximate two week intervals (weeks 0, 2, 4, 6, 9, 12). Severity of depression at any visit was assessed using clinician-rated and self-reported versions of the Quick Inventory of Depressive Symptomatology (QIDS) score (Rush et al., 2003), for which higher values correspond to higher severity. At the end of each level, patients deemed to have an adequate clinical response to that level’s treatment did not move on to future levels, where an adequate response was defined by 12-week clinician-rated QIDS score ≤ 5 (remission) or showing a 50% or greater decrease from the baseline score at the beginning of level 1 (successful reduction). During level 1, all patients were treated with citalopram. Patients continuing to level 2 due to inadequate response were eligible to receive one of up to seven treatment options. We classify these options as either (i) switch: sertraline, bupropion, venlafaxine, or cognitive therapy, or (ii) augment: citalopram plus one of either bupropion, buspirone, or cognitive therapy. Patients assigned to cognitive therapy (alone or augmented with citalopram) were eligible, in the case of inadequate response, to move to a supplementary level 2A and switch to either bupropion or venlafaxine. All patients without adequate response at level 2 (or 2A, if applicable) continued to level 3. Level 3 treatments can again be classified as either (i) switch: mirtazepine or nortriptyline or (ii) augment with either: lithium

or triiodothyronine. Patients without adequate clinical response continued to level 4, requiring a switch to either tranlycypromine or mirtazepine combined with venlafaxine. For a complete description see [Rush et al. \(2004\)](#).

To demonstrate formulation of this problem within the framework of Sections 2 and 3, we take level 2A to be part of level 2 and consider only levels 2 and 3 of the study, calling them stages (decision points) 1 and 2, respectively ($K = 2$). Hence, we include in the analysis only the 1260 patients who entered level 2; 330 of these subsequently continued to level 3. Let A_k , $k = 1, 2$, be the treatment assigned at stage k (beginning of level $k + 1$), taking values 0 (augment) or 1 (switch); both options are feasible for all eligible subjects. Let S_{10} denote baseline QIDS score and S_{11} denote the most recent QIDS score at level 1/beginning of level 2, respectively, so that $S_1 = (S_{10}, S_{11})^T$ is information available immediately prior to the first decision. Similarly, let S_2 be the information available immediately prior to decision 2; here, S_2 is the most recent QIDS score at the end of level 2/beginning of level 3. Finally, let T be QIDS score at the end of level 3. Because some patients exhibited adequate response at the end of level 2 and did not progress to level 3, we define the outcome of interest to be $-S_2$ (negative QIDS score at the end of level 2) for patients not moving to level 3 and $-(S_2 + T)/2$ (average of negative QIDS scores at the end of levels 2 and 3) otherwise. Thus, writing $L_0 = \max(5, S_{10}/2)$, $Y = -S_2 I(S_2 \leq L_0) - (S_2 + T) I(S_2 > L_0)/2$, the cumulative average negative QIDS score. Thus, this demonstrates the case where outcome is a function of accrued information over the sequence of decisions.

It is straightforward to deduce from (14) that $Q_2(\bar{s}_2, \bar{a}_2) = E(Y | \bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2) = -s_2 \{I(s_2 \leq l_0) + I(s_2 > l_0)/2\} + E(-T | \bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2, S_2 > l_0) I(s_2 > l_0)/2$, so that $V_2(\bar{s}_2, a_1) = -s_2 I(s_2 \leq l_0) + \{-s_2 + U_2(\bar{s}_2, a_1)\} I(s_2 > l_0)/2$, where $U_2(\bar{s}_2, a_1) = \max_{a_2} E(-T | \bar{S}_2 = \bar{s}_2, \bar{A}_1 = \bar{a}_1, A_2 = a_2, S_2 > l_0)$. Thus, from (17),

$$Q_1(s_1, a_1) = E[-S_2 I(S_2 \leq l_0) + \{-S_2 + U_2(\bar{s}_2, a_1)\} I(S_2 > l_0) / 2 | S_1 = s_a, A_1 = a_1].$$

We describe implementation for Q -learning. At the second decision point, we must posit a model for $Q_2(\bar{s}_2, \bar{a}_2)$. From the form of $Q_2(\bar{s}_2, \bar{a}_2)$, we need only specify a model for $E(-T | \bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2, S_2 > l_0)$; given the form of the conditioning set, this may be carried out using only the data from patients moving to

level 3. Based on exploratory analysis, defining s_{22} to be the slope of QIDS score over level 2 based on s_{11} and s_2 , we took this model to be of the form $\beta_{20} + \beta_{21}s_2 + \beta_{22}s_{22} + \psi_{20}a_2$, so that the posited Q -function is

$$Q_2(\bar{s}_2, \bar{a}_2; \xi_2) = -s_2\{I(s_2 \leq l_0) + I(s_2 > l_0)/2\} + I(s_2 > l_0)(\beta_{20} + \beta_{21}s_2 + \beta_{22}s_{22} + \psi_{20}a_2)/2, \quad (33)$$

$\xi_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \psi_{20})^T$. Under (33), $V_2(\bar{s}_2, a_1; \xi_2) = -s_2\{I(s_2 \leq l_0) + I(s_2 > l_0)/2\} + I(s_2 > l_0)\{\beta_{20} + \beta_{21}s_2 + \beta_{22}s_{22} + \psi_{20}I(\psi_{20} > 0)\}/2$, and the “responses” $\tilde{V}_{2,i}$ for use in (24) may then be formed by substituting the estimate for ξ_2 . Based on exploratory analysis, we took the posited Q -function at the first stage to be $Q_1(s_1, a_1; \xi_1) = \beta_{10} + \beta_{11}s_{11} + \beta_{12}s_{12} + a_1(\psi_{10} + \psi_{11}s_{12})$, where s_{12} is the slope of QIDS score over level 1 based on s_{10} and s_{11} ; and $\xi_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \psi_{10}, \psi_{11})^T$. For A -learning, we posited models for the functions $h_k(\bar{s}_k, \bar{a}_{k-1})$ and $C_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, 2$, in the obvious way analogous to the models above, and we took the propensity models to be of the form $\pi_2(\bar{s}_2, a_1; \phi_2) = \text{expit}(\phi_{20} + \phi_{21}s_2 + \phi_{22}s_{22} + \phi_{23}a_1)$ and $\pi_1(s_1; \phi_1) = \text{expit}(\phi_{10} + \phi_{11}s_{11} + \phi_{12}s_{12})$.

The results are presented in Table 2. At the first stage, Q -learning suggests a treatment switch for those with level 1 QIDS slope greater than -0.97 (obtained by solving $1.12 + 1.15S_{12} = 0$); A -learning assigns a treatment switch for those with QIDS slope during level 1 greater than -1.07. At the second stage (level 3), the results suggest that all patients should switch treatment and not augment their existing treatments.

7 Discussion

We have provided a self-contained account of Q - and A -learning methods for estimating optimal dynamic treatment regimes, including a detailed discussion of the underlying statistical framework in which these methods may be formalized and of their relative merits. Our simulation studies confirm that, while A -learning may be inefficient relative to Q -learning in estimating parameters that define the optimal regime when the Q -functions required for the latter are correctly specified, A -learning may offer robustness to such misspecification. Nonetheless, Q -learning may have practical advantages in that it involves modeling tasks familiar to most data analysts, allowing the use of standard diagnostic tools. On the other hand,

Table 2: *STAR*D data analysis results. Asterisks indicate evidence at level of significance 0.05 that the parameter is non-zero*

	Q-learning			A-learning				
Parameter	Estimate	95% CI		Estimate	95% CI			
Stage 2								
β_{20}	-1.46	(-3.47 , 0.55)		-1.47	(-3.49 , 0.54)			
β_{21}	-0.75	(-0.88 , -0.61)		*	-0.75	(-0.88 , -0.61)	*	
β_{22}	1.17	(0.52 , 1.81)		*	1.17	(0.52 , 1.81)	*	
ψ_{20}	1.10	(0.02 , 2.19)		*	1.12	(0.03 , 2.22)	*	
Stage 1								
β_{10}	-1.12	(-2.22 , -0.03)		*	-0.90	(-2.03 , 0.22)		
β_{11}	-0.58	(-0.65 , -0.51)		*	-0.59	(-0.66 , -0.52)		*
β_{12}	0.01	(-0.42 , 0.45)			0.11	(-0.34 , 0.57)		
ψ_{10}	1.12	(0.43 , 1.80)		*	0.90	(0.17 , 1.64)		*
ψ_{12}	1.15	(0.20 , 2.10)		*	0.84	(-0.24 , 1.92)		

A-learning may be preferred in settings where it is expected that the form of the decision rules defining the optimal regime is not overly complex. However, A-learning increases in complexity with more than two treatment options at each stage, which may limit its appeal. Interestingly, our simulations demonstrate that inefficiency and bias in estimation of parameters defining the optimal regime does not necessarily translate into degradation of performance of the estimated regime for either method.

There remain many unresolved issues in estimation of optimal treatment regimes using these and other methods. Approaches to address the challenges of high-dimensional information and large numbers of decision points are required. Existing methods for model selection focusing on minimization of prediction error may not be best for developing models optimal for decision-making. Formal inference procedures for evaluating the uncertainty associated with estimation of the optimal regime are challenging due to the nonsmooth nature of decision rules, which in turn leads to nonregularity of the parameter estimators; see [Chakraborty et al. \(2010\)](#), [Laber et al. \(2010\)](#), [Song et al. \(2010\)](#), and [Laber and Murphy \(2011\)](#).

We have discussed sequential decision-making in the context of personalized medicine, but many other applications of these methods exist where, at one or more times in an evolving process, an action must be taken from among a set of plausible actions. Indeed, Q-learning was originally proposed in the computer science literature with these more general problems in mind; see [Shortreed et al. \(2010\)](#) for discussion.

Appendix

A.1 Demonstration That (5)–(8) Define an Optimal Regime

For $k = 1, \dots, K$ and any $d \in \mathcal{D}$, define the random variables $\alpha_k\{\bar{S}_k^*(\bar{d}_{k-1})\}$ such that

$$\alpha_k\{\bar{S}_k^*(\bar{d}_{k-1})\}(\omega) = V_k^{(1)}(\bar{s}_k, \bar{u}_{k-1}) \quad (\text{A.1})$$

for any $\omega \in \Omega$, where $(\bar{s}_k, \bar{u}_{k-1})$ are defined by (3). We now argue that $d^{(1)\text{opt}}$ is an optimal regime; i.e., $d^{(1)\text{opt}}$ satisfies (4). We first show that, for any $d \in \mathcal{D}$,

$$\begin{aligned} \mathbb{E}\{Y^*(d)|S_1 = s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-1})\} &\leq \mathbb{E}\{Y^*(\bar{d}_{K-1}, d_K^{(1)\text{opt}})|S_1 = s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-1})\} \\ &= \alpha_K\{s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-1})\}. \end{aligned} \quad (\text{A.2})$$

This follows because, for the set in Ω where $\{S_2^*(d_1) = s_2, \dots, S_K^*(\bar{d}_{K-1}) = s_K\}$, the left- and right-hand sides of the first line of (A.2) are equal to

$$\begin{aligned} \mathbb{E}\{Y^*(d)|\bar{S}_K^*(\bar{d}_{K-1}) = \bar{s}_K\} &= \mathbb{E}\{Y^*(\bar{u}_{K-1}, u_K)|\bar{S}_K^*(\bar{u}_{K-1}) = \bar{s}_K\}, \quad (\text{A.3}) \\ \mathbb{E}\{Y^*(\bar{d}_{K-1}, d_K^{(1)\text{opt}})|\bar{S}_K^*(\bar{d}_{K-1}) = \bar{s}_K\} &= \mathbb{E}[Y^*\{\bar{u}_{K-1}, d_K^{(1)\text{opt}}(\bar{s}_K, \bar{u}_{K-1})\}|\bar{S}_K^*(\bar{u}_{K-1}) = \bar{s}_K] \quad (\text{A.4}) \end{aligned}$$

respectively. By the definition of $d_K^{(1)\text{opt}}$ in (5), (A.4) is greater than or equal to (A.3), and, by the definition of $V_K^{(1)}$ in (6), (A.4) equals $V_K^{(1)}(\bar{s}_K, \bar{u}_{K-1})$. Because these results hold for sets $\{S_2^*(d_1) = s_2, \dots, S_K^*(\bar{d}_{K-1}) = s_K\}$ for any (s_2, \dots, s_K) , and by the definition of α_K in (A.1), (A.2) holds. Taking conditional expectations given $S_1 = s_1$ yields

$$\begin{aligned} \mathbb{E}\{Y^*(d)|S_1 = s_1\} &\leq \mathbb{E}\{Y^*(\bar{d}_{K-1}, d_K^{(1)\text{opt}})|S_1 = s_1\} \\ &= \mathbb{E}[\alpha_K\{s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-1})\}|S_1 = s_1]. \end{aligned} \quad (\text{A.5})$$

The equality in (A.5) holds for any $\bar{d}_{K-1} = (d_1, \dots, d_{K-1})$, hence it must hold for $(d_1, \dots, d_{K-2}, d_{K-1}^{(1)\text{opt}})$.

Thus, we also have that

$$\begin{aligned} & \mathbb{E}\{Y^*(\bar{d}_{K-2}, d_{K-1}^{(1)\text{opt}}, d_K^{(1)\text{opt}}) | S_1 = s_1\} \\ &= \mathbb{E}[\alpha_K \{S_1, S_2^*(d_1), \dots, S_{K-1}^*(\bar{d}_{K-2}), S_K^*(\bar{d}_{K-2}, d_{K-1}^{(1)\text{opt}})\} | S_1 = s_1]. \end{aligned} \quad (\text{A.6})$$

Similarly, for any $k = K-1, \dots, 1$, we can show that $\mathbb{E}[\alpha_{k+1} \{S_1, S_2^*(d_1), \dots, S_{k+1}^*(\bar{d}_k)\} | S_1 = s_1, S_2^*(d_1), \dots, S_k^*(\bar{d}_{k-1})] \leq \mathbb{E}[\alpha_{k+1} \{S_1, S_2^*(d_1), \dots, S_{k+1}^*(\bar{d}_{k-1}, d_k^{(1)\text{opt}})\} | S_1 = s_1, S_2^*(d_1), \dots, S_k^*(\bar{d}_{k-1})] = \alpha_k \{s_1, S_2^*(d_1), \dots, S_k^*(\bar{d}_{k-1})\}$, which implies for $k = K-1, \dots, 1$,

$$\begin{aligned} & \mathbb{E}[\alpha_{k+1} \{s_1, S_2^*(d_1), \dots, S_{k+1}^*(\bar{d}_k)\} | S_1 = s_1] \leq \mathbb{E}[\alpha_{k+1} \{s_1, S_2^*(d_1), \dots, S_{k+1}^*(\bar{d}_{k-1}, d_k^{(1)\text{opt}})\} | S_1 = s_1] \\ &= \mathbb{E}[\alpha_k \{s_1, S_2^*(d_1), \dots, S_k^*(\bar{d}_{k-1})\} | S_1 = s_1] \end{aligned} \quad (\text{A.7})$$

Using (A.5) and (A.7) with $k = K-1$, we thus have

$$\begin{aligned} & \mathbb{E}\{Y^*(d) | S_1 = s_1\} \leq \mathbb{E}\{Y^*(\bar{d}_{K-1}, d_K^{(1)\text{opt}}) | S_1 = s_1\} = \mathbb{E}[\alpha_K \{s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-1})\} | S_1 = s_1] \\ & \leq \mathbb{E}[\alpha_K \{s_1, S_2^*(d_1), \dots, S_K^*(\bar{d}_{K-2}, d_{K-1}^{(1)\text{opt}})\} | S_1 = s_1] \\ &= \mathbb{E}[\alpha_{K-1} \{s_1, S_2^*(d_1), \dots, S_{K-1}^*(\bar{d}_{K-2})\} | S_1 = s_1] \end{aligned} \quad (\text{A.8})$$

Because of (A.6), the term in (A.8) is equal to $\mathbb{E}\{Y^*(\bar{d}_{K-2}, d_{K-1}^{(1)\text{opt}}, d_K^{(1)\text{opt}}) | S_1 = s_1\}$. Hence,

$$\begin{aligned} & \mathbb{E}\{Y^*(d) | S_1 = s_1\} \leq \mathbb{E}\{Y^*(\bar{d}_{K-1}, d_K^{(1)\text{opt}}) | S_1 = s_1\} \leq \mathbb{E}\{Y^*(\bar{d}_{K-2}, d_{K-1}^{(1)\text{opt}}, d_K^{(1)\text{opt}}) | S_1 = s_1\} \\ &= \mathbb{E}[\alpha_{K-1} \{s_1, S_2^*(d_1), \dots, S_{K-1}^*(\bar{d}_{K-2})\} | S_1 = s_1]. \end{aligned} \quad (\text{A.9})$$

Again, because \bar{d}_{K-2} is arbitrary, if we replace it by $(\bar{d}_{K-3}, d_{K-2}^{(1)\text{opt}})$, the equality in (A.9) implies

$$\mathbb{E}\{Y^*(\bar{d}_{K-3}, d_{K-2}^{(1)\text{opt}}) | S_1 = s_1\} = \mathbb{E}[\alpha_{K-1} \{s_1, S_2^*(d_1), \dots, S_{K-1}^*(\bar{d}_{K-3}, d_{K-2}^{(1)\text{opt}})\} | S_1 = s_1], \quad (\text{A.10})$$

where, for any d , $\underline{d}_k = (d_k, \dots, d_K)$. Using (A.7) with $k = K - 2$, (A.9), and (A.10), we obtain

$$\begin{aligned} \mathbb{E}\{Y^*(\bar{d}_{K-2}, \underline{d}_{K-1}^{(1)\text{opt}})|S_1 = s_1\} &= \mathbb{E}[\alpha_{K-1}\{s_1, S_2^*(d_1), \dots, S_{K-1}^*(\bar{d}_{K-2})\}|S_1 = s_1] \\ &\leq \mathbb{E}[\alpha_{K-1}\{s_1, S_2^*(d_1), \dots, S_{K-1}^*(\bar{d}_{K-3}, d_{K-2}^{(1)\text{opt}})|S_1 = s_1\} = \mathbb{E}\{Y^*(\bar{d}_{K-3}, \underline{d}_{K-2}^{(1)\text{opt}})\}|S_1 = s_1\} \\ &= \mathbb{E}[\alpha_{K-2}\{s_1, S_2^*(d_1), \dots, S_{K-2}^*(\bar{d}_{K-3})\}|S_1 = s_1]. \end{aligned}$$

Continuing in this fashion, we may conclude that, for any $d \in \mathcal{D}$,

$$\mathbb{E}\{Y^*(d)|S_1 = s_1\} \leq \dots \leq \mathbb{E}\{Y^*(\bar{d}_{k-1}, \underline{d}_k^{(1)\text{opt}})|S_1 = s_1\} \leq \dots \leq \mathbb{E}\{Y^*(d^{(1)\text{opt}})|S_1 = s_1\},$$

showing that $d^{(1)\text{opt}}$ defined in (5) and (7) is an optimal regime satisfying (4).

A.2 Demonstration of Correspondence in (20)–(22) Under Assumptions in Section 2

We first consider the case $\ell = 1$. We make the positivity assumption that, for any $(\bar{s}_k, \bar{a}_{k-1})$ for which $\text{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) > 0$, $\text{pr}(A_k = a_k|\bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}) > 0$ if and only if $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, \dots, K$. This ensures that the observed data contain information on the treatments involved in the class of feasible regimes under consideration. We have $\Gamma_k^{(1)} = \Gamma_k$ by definition, so we need only demonstrate (21) and (22). We must show that, for any $(\bar{s}_k, \bar{a}_{k-1}) \in \Gamma_k$ and $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, \dots, K$,

$$\text{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0, \tag{A.11}$$

$$\text{pr}(S_{k+1} = s_{k+1}|\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) = \text{pr}\{S_{k+1}^*(\bar{a}_k) = s_{k+1}|\bar{S}_k = \bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}\}, \tag{A.12}$$

$$= \text{pr}\{S_{k+1}^*(\bar{a}_k) = s_{k+1}|\bar{S}_j = \bar{s}_j, \bar{A}_{j-1} = \bar{a}_{j-1}, S_{j+1}^*(\bar{a}_j) = s_{j+1}, \dots, S_k^*(\bar{a}_{k-1}) = s_k\}, \tag{A.13}$$

for $j = 1, \dots, k$, where we define (A.13) with $j = k$ to be the same as the expression on the right-hand side of (A.12) and take $S_{k+1} = Y$ and $S_{k+1}^*(\bar{a}_k) = Y^*(\bar{a}_k)$.

Assume for the moment that (A.11) is true. We now demonstrate (A.12) and (A.13). For any fixed k , by the consistency assumption, the left-hand expression in (A.12) is equal to $\text{pr}\{S_{k+1}^*(\bar{a}_k) = s_{k+1}|\bar{S}_k =$

$\bar{s}_k, \bar{A}_{k-1} = \bar{a}_{k-1}, A_k = a_k\}$. It follows by the sequential randomization assumption, which implies $A_k \perp\!\!\!\perp S_{k+1}^*(\bar{a}_k) | \bar{S}_k, \bar{A}_{k-1}$, that this is equal to the right-hand side of (A.12). The equality in (A.13) follows by induction. Specifically, treating the right-hand side of (A.12) as (A.13) with $j = k$, the equality follows if we can show that (A.13) being true for a given j implies that it is also true for $j - 1$. For a given $j = 2, \dots, k$, by the consistency assumption, (A.13) is equal to $\text{pr}\{S_{k+1}^*(\bar{a}_k) = s_{k+1} | \bar{S}_{j-1} = \bar{s}_{j-1}, \bar{A}_{j-2} = \bar{a}_{j-2}, A_{j-1} = a_{j-1}, S_j^*(\bar{a}_j) = s_j, \dots, S_k^*(\bar{a}_{k-1}) = s_k\}$. By the sequential randomization assumption, $A_{j-1} \perp\!\!\!\perp \{S_j^*(\bar{a}_j), \dots, S_{k+1}^*(\bar{a}_k)\} | \bar{S}_{j-1}, \bar{A}_{j-2}$, so that this expression is equal to $\text{pr}\{S_{k+1}^*(\bar{a}_k) = s_{k+1} | \bar{S}_{j-1} = \bar{s}_{j-1}, \bar{A}_{j-2} = \bar{a}_{j-2}, S_j^*(\bar{a}_j) = s_j, \dots, S_k^*(\bar{a}_{k-1}) = s_k\}$, which is (A.13) for $j - 1$. Note, then, that this implies that the conditional densities in (A.13), which are j -dependent, are the same as those on the left-hand side of (A.12), which are not.

We now prove (A.11) by induction. Assume we have shown that $\text{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0$. Then we must show that $\text{pr}(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_{k+1} = \bar{a}_{k+1}) > 0$. If $\text{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0$, then

$$\text{pr}(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k) = \text{pr}(S_{k+1} = s_{k+1} | \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) \text{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k). \quad (\text{A.14})$$

But we have shown above that if (A.11) is true; i.e., $\text{pr}(\bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) > 0$, then (A.12) and (A.13) are equal for all j and in particular $\text{pr}(S_{k+1} = s_{k+1} | \bar{S}_k = \bar{s}_k, \bar{A}_k = \bar{a}_k) = \text{pr}\{S_{k+1}^*(\bar{a}_k) = s_{k+1} | \bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k\}$. Because $(\bar{s}_{k+1}, \bar{a}_k) \in \Gamma_{k+1}$, then by condition (ii) of (2) defining Γ_{k+1} , $\text{pr}\{S_{k+1}^*(\bar{a}_k) = s_{k+1} | \bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k\} > 0$ because of (A.14). Now $\text{pr}(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_{k+1} = \bar{a}_{k+1}) = \text{pr}(A_{k+1} = a_{k+1} | \bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k) \text{pr}(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k)$; however, because $a_{k+1} \in \Psi_k(\bar{s}_{k+1}, \bar{a}_k)$ and by the positivity assumption, $\text{pr}(A_{k+1} = a_{k+1} | \bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_k = \bar{a}_k) > 0$ and hence $\text{pr}(\bar{S}_{k+1} = \bar{s}_{k+1}, \bar{A}_{k+1} = \bar{a}_{k+1}) > 0$. The proof is complete by noting that $\text{pr}(S_1 = s_1, A_1 = a_1) = \text{pr}(A_1 = a_1 | S_1 = s_1) \text{pr}(S_1 = s_1)$, where $\text{pr}(A_1 = a_1 | S_1 = s_1) > 0$ for $a_1 \in \Psi(s_1)$ by the positivity assumption.

To demonstrate (21) and (22) for $\ell = 1$, consider first the definitions of $d_K^{(1)\text{opt}}(\bar{s}_K, \bar{a}_{K-1})$ and $V_K^{(1)}(\bar{s}_K, \bar{a}_{K-1})$ given in (5) and (6). These quantities involve the conditional expectation of the potential outcome $Y^*(\bar{a}_K)$ given $\bar{S}_K^*(\bar{a}_{K-1})$, which by (A.12)-(A.13) is the same as the conditional expectation of Y given $\{\bar{S}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K\}$. Thus, $d_K^{(1)\text{opt}}(\bar{s}_K, \bar{a}_{K-1})$ and $V_K^{(1)}(\bar{s}_K, \bar{a}_{K-1})$ are the same as $d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1})$ and $V_K(\bar{s}_K, \bar{a}_{K-1})$ defined in (15) and (16). Next, from (7) and (8), $d_{K-1}^{(1)\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2}) = \arg \max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} \mathbb{E}[V_K^{(1)}\{\bar{s}_K$

\bar{s}_{K-1}]. This involves the conditional expectation of $V_K^{(1)}$, a function of $S_K^*(\bar{a}_{K-1})$, given $\bar{S}_{K-1}^*(\bar{a}_{K-2}) = \bar{s}_{K-1}$. Again, by (A.12)-(A.13), this is the same as the conditional expectation of the function $V_K^{(1)}$ of S_K given $\{\bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1}\}$. Because we have already shown that $V_K^{(1)}$ is the same as V_K , this implies that $d_{K-1}^{(1)\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2})$ is given by

$$\arg \max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} \mathbb{E}\{V_K(\bar{s}_{K-1}, S_K, \bar{a}_{K-2}, a_{K-1}) | \bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = (\bar{a}_{K-2}, a_{K-2})\},$$

which is the same as $d_{K-1}^{\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2})$ given by (18) with $k = K - 1$. The argument continues in a backward iterative fashion for $k = K - 2, \dots, 1$.

Now consider $\ell > 1$. The sets $\mathcal{V}_{\ell,k}$, $\ell = 1, \dots, K$, $k = \ell, \dots, K$, representing events of the form $\{\bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}, S_{\ell+1}^*(\bar{a}_\ell) = s_{\ell+1}, \dots, S_k^*(\bar{a}_{k-1}) = s_k\}$, involved in the definitions of $\Gamma_k^{(\ell)}$ and (10)-(13), depend on the random variables $\bar{S}_k^{(P)}, \bar{A}_{k-1}^{(P)}$ for $k = \ell, \dots, K$, which characterize how treatment assignment and covariate history arise in the population under routine practice. To demonstrate (20)-(22), in addition to those on the observed random variables given above, we also require sequential randomization and positivity assumptions on the “population” random variables; namely, that $A_k^{(P)} \perp\!\!\!\perp W | \bar{S}_k^{(P)}, \bar{A}_{k-1}^{(P)}$, $k = 1, \dots, K$; and, for any $(\bar{s}_k, \bar{a}_{k-1})$ for which $\text{pr}(\bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_{k-1}^{(P)} = \bar{a}_{k-1}) > 0$, $\text{pr}(A_k^{(P)} = a_k | \bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_{k-1}^{(P)} = \bar{a}_{k-1}) > 0$ if and only if $a_k \in \Psi_k(\bar{s}_k, \bar{a}_{k-1})$, $k = 1, \dots, K$. If the observed data are from an observational study where S_k, A_k are the same as $S_k^{(P)}, A_k^{(P)}$, these assumptions are equivalent to those on the observed data. For data from a SMART, however, more consideration is required. If the treatment options considered in the trial are restricted relative to those available in practice, then an estimated optimal regime based on the observed data may not be applicable to patients who present at the ℓ th decision with treatment histories involving options not considered in the trial for $\ell > 1$. The positivity assumption here rules out such patients from consideration. The sequential randomization assumption holds for observed data by design for a SMART. However, whether or not it holds in the population, as we require here, depends on whether or not the covariate information collected in the trial contains the information used by patients and their providers to make treatment decisions in routine practice. If this is not the case, then the estimated optimal regime based on the trial data is still applicable to patients who present prior to the first decision, $\ell = 1$, but may not lead to optimal decision-making for patients presenting at subsequent

decision points because the sequential randomization assumption at the population level may no longer hold.

Under these assumptions, it follows by an argument analogous to that above that (A.11)–(A.13) hold with the random variables S_k, A_k replaced by $S_k^{(P)}, A_k^{(P)}$, $k = 1, \dots, K$; namely

$$\text{pr}(\bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_k^{(P)} = \bar{a}_k) > 0, \quad (\text{A.15})$$

$$\text{pr}(S_{k+1}^{(P)} = s_{k+1} | \bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_k^{(P)} = \bar{a}_k) = \text{pr}\{S_{k+1}^*(\bar{a}_k) = s_{k+1} | \bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_{k-1}^{(P)} = \bar{a}_{k-1}\}, \quad (\text{A.16})$$

$$= \text{pr}\{S_{k+1}^*(\bar{a}_k) = s_{k+1} | \bar{S}_j^{(P)} = \bar{s}_j, \bar{A}_{j-1}^{(P)} = \bar{a}_{j-1}, S_{j+1}^*(\bar{a}_j) = s_{j+1}, \dots, S_k^*(\bar{a}_{k-1}) = s_k\}, \quad (\text{A.17})$$

for $j = 1, \dots, k$. We may then show that (20) holds as follows. Inspection of Γ_k and $\Gamma_k^{(\ell)}$ shows both sets involve the same condition (i). Accordingly, we need only demonstrate that, if condition (ii) in $\Gamma_k^{(\ell)}$ holds, then so does (ii) in Γ_k , and vice versa. Condition (ii) in $\Gamma_k^{(\ell)}$ states that $\text{pr}(\mathcal{V}_{\ell,k}) > 0$. Because the set $\mathcal{V}_{\ell,k} \subseteq \{\bar{S}_k^*(\bar{a}_{k-1}) = \bar{s}_k\}$, condition (ii) in Γ_k follows immediately. In the converse direction, if (ii) of Γ_k holds, then (A.15) holds. Because the set $\{\bar{S}_k^{(P)} = \bar{s}_k, \bar{A}_k^{(P)} = \bar{a}_k\} \subseteq \mathcal{V}_{\ell,k}$, $\text{pr}(\mathcal{V}_{\ell,k}) > 0$, which is (ii) of $\Gamma_k^{(\ell)}$.

Now (21) and (22) follow by an argument similar to that for $\ell = 1$. First, we argue that, for any fixed $k = 1, \dots, K$, the probabilities in (A.12) and (A.13) are the same as those in (A.16) and (A.17) for all $j = 1, \dots, k$. This follows because (A.13) with $j = 1$ is equal to (A.17) with $j = 1$. We may now use this to show the result. Consider the definitions of $d_K^{(\ell)\text{opt}}(\bar{s}_K, \bar{a}_{K-1})$ and $V_K^{(\ell)}(\bar{s}_K, \bar{a}_{K-1})$ given in (10) and (11). These quantities involve the conditional expectation of the potential outcome $Y^*(\bar{a}_K)$ given $\{\bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}, S_{\ell+1}^*(\bar{a}_\ell) = s_{\ell+1}, \dots, S_K^*(\bar{a}_{K-1}) = \bar{a}_{K-1}\}$. But, because of the above equivalence of (A.12)–(A.13) and (A.16)–(A.17), this is the same as the conditional expectation of Y given $\{\bar{S}_K = \bar{s}_K, \bar{A}_K = \bar{a}_K\}$. Thus, $d_K^{(\ell)\text{opt}}(\bar{s}_K, \bar{a}_{K-1})$ and $V_K^{(\ell)}(\bar{s}_K, \bar{a}_{K-1})$ are the same as $d_K^{\text{opt}}(\bar{s}_K, \bar{a}_{K-1})$ and $V_K(\bar{s}_K, \bar{a}_{K-1})$ defined in (15) and (16), and this is true for all $\ell = 2, \dots, K$. Next, in accordance with (21) and (22), $d_{K-1}^{(\ell)\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2}) = \arg \max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} \mathbb{E}[V_K^{(\ell)}\{\bar{s}_{K-1}, S_K^*(\bar{a}_{K-2}, a_{K-1}), \bar{a}_{K-2}, a_{K-1}\} | \bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}, S_{\ell+1}^*(\bar{a}_\ell) = s_{\ell+1}, \dots, S_{K-1}^*(\bar{a}_{K-2}) = s_{K-1}]$. Note that this involves the conditional expectation of the function $V_K^{(\ell)}$ of $S_K^*(\bar{a}_{K-1})$ given $\bar{S}_\ell^{(P)} = \bar{s}_\ell, \bar{A}_{\ell-1}^{(P)} = \bar{a}_{\ell-1}, S_{\ell+1}^*(\bar{a}_\ell) = s_{\ell+1}, \dots, S_{K-1}^*(\bar{a}_{K-2}) = s_{K-1}$. Again, this is the same as the conditional expectation of the function $V_K^{(\ell)}$ of S_K given $\{\bar{S}_K = \bar{s}_K, \bar{A}_{K-1} = \bar{a}_{K-1}\}$. Because we

have shown that $V_K^{(\ell)}$ is independent of ℓ and equal to V_K , this implies that

$$d_{K-1}^{(\ell)\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2}) = \arg \max_{a_{K-1} \in \Psi_{K-1}(\bar{s}_{K-1}, \bar{a}_{K-2})} \mathbb{E}\{V_K(\bar{s}_{K-1}, S_K, \bar{a}_{K-2}, a_{K-1}) | \bar{S}_K = \bar{s}_K, \bar{A}_{K-1}\},$$

which is the same as $d_{K-1}^{\text{opt}}(\bar{s}_{K-1}, \bar{a}_{K-2})$ given by (18) with $k = K - 1$. The argument continues in an backward iterative fashion for $k = K - 2, \dots, 1$.

A.3 Justification for \tilde{V}_{ki} in A-learning

We wish to show that

$$E \left(V_{k+1}(\bar{S}_{k+1}, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1}) [I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k] \middle| \bar{S}_k, \bar{A}_{k-1} \right) = V_k(\bar{S}_k, \bar{A}_{k-1}). \quad (\text{A.18})$$

Defining $\Gamma(\bar{S}_{k+1}, \bar{A}_k) = V_{k+1}(\bar{S}_{k+1}, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1}) [I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k]$, we may write (A.18) as

$$\mathbb{E}[\mathbb{E}\{\Gamma(\bar{S}_{k+1}, \bar{A}_k) | \bar{S}_k, \bar{A}_{k-1}\} | \bar{S}_k, \bar{A}_{k-1}]. \quad (\text{A.19})$$

The inner expectation in (A.19) may be seen to be equal to

$$\begin{aligned} & \mathbb{E}\{V_{k+1}(\bar{S}_{k+1}, \bar{A}_k) | \bar{S}_k, \bar{A}_{k-1}\} + C_k(\bar{S}_k, \bar{A}_{k-1}) [I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k] \\ &= Q_k(\bar{S}_k, \bar{A}_k) + C_k(\bar{S}_k, \bar{A}_{k-1}) [I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} - A_k]. \end{aligned}$$

Substituting $Q_k(\bar{S}_k, \bar{A}_k) = h_k(\bar{S}_k, \bar{A}_{k-1}) + A_k C_k(\bar{S}_k, \bar{A}_{k-1})$, $h_k(\bar{S}_k, \bar{A}_{k-1}) = Q_k(\bar{S}_k, \bar{A}_{k-1}, 0)$, we obtain $\mathbb{E}\{\Gamma(\bar{S}_{k+1}, \bar{A}_k) | \bar{S}_k, \bar{A}_{k-1}\} = h_k(\bar{S}_k, \bar{A}_{k-1}) + C_k(\bar{S}_k, \bar{A}_{k-1}) I\{C_k(\bar{S}_k, \bar{A}_{k-1}) > 0\} = V_k(\bar{S}_k, \bar{A}_{k-1})$. Substituting this in (A.19) yields the result.

A.4 Demonstration of Equivalence of Q - and A -learning in a Special Case

We take $K = 1$ and let $\text{pr}(A_1 = 1|S_1 = s_1) = \pi$. Consider the A -learning estimating equations (28) with $k = 1$, and take $\lambda_1(s_1; \psi_1) = \partial/\partial\psi_1 C_1(s_1; \psi_1)$. Then the equations become

$$\sum_{i=1}^n \frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1} (A_{1i} - \pi) \{Y_i - A_{1i} C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0,$$

$$\sum_{i=1}^n \frac{\partial h_1(S_{1i}; \beta_1)}{\partial \beta_1} \{Y_i - A_{1i} C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0.$$

Likewise, under these conditions, taking $Q_1(s_1, a_1) = a_1 C_1(s_1; \psi_1) + h_1(s_1; \beta_1)$, the Q -learning equation is

$$\sum_{i=1}^n \frac{\partial Q_1(S_{1i}, A_{1i}; \xi_1)}{\partial \xi_1} \{Y_i - A_{1i} C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0,$$

where, with $\xi_1 = (\psi_1^T, \beta_1^T)^T$,

$$\frac{\partial Q_1(S_{1i}, A_{1i}; \xi_1)}{\partial \xi_1} = \begin{pmatrix} A_{1i} \frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1} \\ \frac{\partial h_1(S_{1i}; \beta_1)}{\partial \beta_1} \end{pmatrix}.$$

Thus note that, with $C_1(s_1; \psi_1)$ and $h_1(s_1; \beta_1)$ linear in functions of S_1 , as long as terms of the form in $C_1(s_1; \psi_1)$ are contained in those in $h_1(s_1; \beta_1)$, the Q - and A -learning estimating equations are identical, as then

$$\sum_{i=1}^n \frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1} \{Y_i - A_{1i} C_1(S_{1i}; \psi_1) - h_1(S_{1i}; \beta_1)\} = 0.$$

For example, if $C_1(s_1; \psi_1) = \psi_{10} + s_1^T \psi_{11}$ and $h_1(s_1; \beta_1) = \beta_{10} + s_1^T \beta_{11}$, then note that

$$\frac{\partial C_1(S_{1i}; \psi_1)}{\partial \psi_1} = \frac{\partial h_1(S_{1i}; \beta_1)}{\partial \beta_1} = \begin{pmatrix} 1 \\ S_{1i} \end{pmatrix},$$

and the result is immediate.

A.5 Example of Incompatibility of Q -function Models

To show (30), noting $\mathcal{H}_2 = (1, s_1, a_1, s_2)^T = (\mathcal{K}_1^T, s_2)^T$, we have

$$\begin{aligned} \mathbb{E}\{V_2(s_1 S_2, a_1; \xi_2) | S_1 = s_1, A_1 = a_1\} &= \mathcal{K}_1^T \beta_{21} + \beta_{22} \mathbb{E}(S_2 | S_1 = s_1, A_1 = a_1) \\ &+ (\mathcal{K}_1^T \psi_{21}) \mathbb{E}\{I(\mathcal{K}_1^T \psi_{21} + S_2 \psi_{22} > 0) | S_1 = s_1, A_1 = a_1\} \\ &+ \psi_{22} \mathbb{E}\{S_2 I(\mathcal{K}_1^T \psi_{21} + S_2 \psi_{22} > 0) | S_1 = s_1, A_1 = a_1\}. \end{aligned}$$

Taking $\psi_{22} > 0$, we also have $I(\mathcal{K}_1^T \psi_{21} + S_2 \psi_{22} > 0) = I(S_2 > -\mathcal{K}_1^T \psi_{21}/\psi_{22})$, from which it follows that $\mathbb{E}\{I(\mathcal{K}_1^T \psi_{21} + S_2 \psi_{22} > 0) | S_1 = s_1, A_1 = a_1\} = 1 - \Phi\{(-\mathcal{K}_1^T \psi_{21}/\psi_{22} - \mathcal{K}_1^T \gamma)/\sigma\} = 1 - \Phi(\eta)$ for $\eta = -\mathcal{K}_1^T (\psi_{21}/\psi_{22} + \gamma)/\sigma$. Similarly, $\mathbb{E}\{S_2 I(\mathcal{K}_1^T \psi_{21} + S_2 \psi_{22} > 0) | S_1 = s_1, A_1 = a_1\} = \mathbb{E}\{S_2 I(S_2 > -\mathcal{K}_1^T \psi_{21}/\psi_{22}) | S_1 = s_1, A_1 = a_1\}$. It is straightforward to deduce that this is equal to $\int_{\eta}^{\infty} (\sigma t + \mathcal{K}_1^T \gamma) \varphi(t) dt = \sigma \varphi(\eta) + (\mathcal{K}_1^T \gamma) \{1 - \Phi(\eta)\}$. Using $\mathbb{E}(S_2 | S_1 = s_1, A_1 = a_1) = \mathcal{K}_1^T \gamma$ and combining yields (30).

A.6 Calculation of $\mathbb{E}\{H(\hat{d}^{\text{opt}})\}$ and $R(\hat{d}^{\text{opt}})$

Calculation for $K = 1$. We consider the generative data model in Section 5.1 and treatment regimes of the form $d(s_1) = d_1(s_1) = I(\psi_{10} + \psi_{11} s_1 > 0)$ for arbitrary ψ_{10}, ψ_{11} . It is possible to derive analytically $H(d) = \mathbb{E}\{Y^*(d)\}$ in this case. Under the generative data model, $\mathbb{E}\{Y^*(d)\} = \mathbb{E}[\mathbb{E}\{Y^*(d) | S_1\}] = \mathbb{E}[\mathbb{E}\{Y | S_1, A_1 = d_1(S_1)\}] = \beta_{10}^0 + \beta_{11}^0 \mathbb{E}(S_1) + \beta_{12}^0 \mathbb{E}(S_1^2) + \mathbb{E}\{I(\psi_{10} + \psi_{11} S_1 > 0)(\psi_{10}^0 + \psi_{11}^0 S_1)\}$, and $S_1 \sim \text{Normal}(0, 1)$. It is straightforward to deduce that $\mathbb{E}\{I(\psi_{10} + \psi_{11} S_1 > 0)\} = \text{pr}(S_1 > -\psi_{10}/\psi_{11})$ or $\text{pr}(S_1 < -\psi_{10}/\psi_{11})$ as $\psi_{11} > 0$ or $\psi_{11} < 0$, which is readily obtained from the standard normal cdf. Likewise, $\mathbb{E}\{S_1 I(\psi_{10} + \psi_{11} S_1 > 0)\} = \mathbb{E}(S_1 | S_1 > -\psi_{10}/\psi_{11}) \text{pr}(S_1 > -\psi_{10}/\psi_{11})$ if $\psi_{11} > 0$ and $\mathbb{E}\{S_1 I(\psi_{10} + \psi_{11} S_1 > 0)\} = \mathbb{E}(S_1 | S_1 < -\psi_{10}/\psi_{11}) \text{pr}(S_1 < -\psi_{10}/\psi_{11})$ if $\psi_{11} < 0$, which are again easily calculated in a manner similar to that in Section A.5. Thus, $\mathbb{E}\{Y^*(d^{\text{opt}})\}$ is obtained by substituting ψ_{10}^0, ψ_{11}^0 in the resulting expression. To approximate $\mathbb{E}\{H(\hat{d}^{\text{opt}})\}$ and hence $R(\hat{d}^{\text{opt}})$ for $\hat{d}^{\text{opt}} = \hat{d}_Q^{\text{opt}}$ or \hat{d}_A^{opt} , we may use Monte Carlo simulation. Specifically, for the b th of B Monte Carlo data sets, substitute the estimates $\hat{\psi}_{10,b}, \hat{\psi}_{11,b}$, say, defining \hat{d}^{opt} for that data set in the expression for $\mathbb{E}\{Y^*(d)\}$, and call the resulting quantity U_b . Then $\mathbb{E}\{H(\hat{d}^{\text{opt}})\}$ is

approximated by $B^{-1} \sum_{b=1}^B U_b$. Combining yields the approximation to $R(\hat{d}^{\text{opt}})$.

Calculation for $K = 2$. The developments are analogous to those above. We consider the generative data model in Section 5.2 and treatment regimes of the form $d = (d_1, d_2)$, where $d_1(s_1) = I(\psi_{10} + \psi_{11}s_1 > 0)$ and $d_2(s_1, s_2, a_1) = I(\psi_{20} + \psi_{21}a_1 + \psi_{22}s_2 > 0)$ for arbitrary $\psi_{10}, \psi_{11}, \psi_{20}, \psi_{21}, \psi_{22}$. Here, $E\{Y^*(d)\} = E\left(E\{E\{Y^*(d)|S_2^*(d), S_1\}|S_1\}\right) = E\left\{E\left(E\{Y|S_2, S_1, A_1 = d_1(S_1), A_2 = d_2\{S_2, S_1, d_1(S_1)\}\}\right)\middle|S_1, A_1 = d_1(S_1)\right\}$. Because S_1 is binary taking values in $\{0, 1\}$, $E\{Y^*(d)\} = E\left(E\{Y|S_2, S_1, A_1 = d_1(0), A_2 = d_2\{S_2, 0, d_1(0)\}\}\middle|S_1 = 0, A_1 = 0\right) + E\left(E\{Y|S_2, S_1, A_1 = d_1(1), A_2 = d_2\{S_2, 1, d_1(1)\}\}\middle|S_1 = 1, A_1 = d_1(1)\right) \text{pr}(S_1 = 1)$. Under the generative model, writing $a_1 = I(\psi_{10} + \psi_{11}s_1 > 0)$ for brevity, these expectations are of the form $E\left(E\{Y|S_2, S_1, A_1 = d_1(s_1), A_2 = d_2\{S_2, s_1, d_1(s_1)\}\}\middle|S_1 = s_1, A_1 = d_1(s_1)\right) = \beta_{20} + \beta_{21}^0 s_1 + \beta_{22}^0 a_1 + \beta_{23}^0 s_1 a_1 + \beta_{24}^0 E\{(S_2|S_1 = s_1, A_1 = d_1(s_1))\} + \beta_{25}^0 E\{S_2^2|S_1 = s_1, A_1 = d_1(s_1)\} + (\psi_{20}^0 + \psi_{21}^0 a_1)E\{I(\psi_{20} + \psi_{21}a_1 + \psi_{22}S_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\} + \psi_{22}^0 E\{S_2 I(\psi_{20} + \psi_{21}a_1 + \psi_{22}S_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\}$, for $s_1 = 0, 1$. In the generative data model, the conditional distribution of S_2 given S_1, A_1 is normal; accordingly, it is straightforward to calculate $E\{S_2|S_1 = s_1, A_1 = d_1(s_1)\}$, $E\{S_2^2|S_1 = s_1, A_1 = d_1(s_1)\}$, $E\{I(\psi_{20} + \psi_{21}a_1 + \psi_{22}S_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\}$, and $E\{S_2 I(\psi_{20} + \psi_{21}a_1 + \psi_{22}S_2 > 0)|S_1 = s_1, A_1 = d_1(s_1)\}$ in a manner analogous to those for the case $K = 1$. Approximation of $E\{H(\hat{d}^{\text{opt}})\}$ and hence $R(\hat{d}^{\text{opt}})$ for $\hat{d}^{\text{opt}} = \hat{d}_Q^{\text{opt}}$ or \hat{d}_A^{opt} may then be carried out as for the case $K = 1$.

Calculation by simulation. When an analytical expression for $H(d) = E\{Y^*(d)\}$ for regimes of a certain form d is not available, $H(d)$ for a fixed d may be approximated by simulation using the g-computation algorithm of Robins (1986). We demonstrate for $K = 2$, so that $d = (d_1, d_2)$; the procedure for $K = 1$ is then immediate. For total number of simulations B , for each $b = 1, \dots, B$, the steps are: (i) Generate s_{1b} from the true distribution of S_1 ; (ii) generate s_{2b} from the true conditional distribution of S_2 given $S_1 = s_{1b}$ and $A_1 = d_1(s_{1b})$; (iii) evaluate the true $E(Y|\bar{S}_2 = \bar{s}_2, \bar{A}_2 = \bar{a}_2)$ at $\bar{s}_2 = \bar{s}_{2b} = (s_{1b}, s_{2b})$ and $\bar{a}_2 = [d_1(s_{1b}), d_2\{\bar{s}_{2b}, d_1(s_{1b})\}]$, and call the resulting value U_b ; and (iv) estimate $H(d) = E\{Y^*(d)\}$ by $B^{-1} \sum_{b=1}^B U_b$. When $d = \hat{d}_Q^{\text{opt}}$ or \hat{d}_A^{opt} , one would follow the above procedure for each Monte Carlo data set. In each of steps (i)–(iii), it is important to recognize that, while \hat{d}_Q^{opt} and \hat{d}_A^{opt} are determined by the estimated ψ , the distributions from which realizations are generated depend on the true β and ψ . The values of $E\{H(\hat{d}_Q^{\text{opt}})\}$ and $E\{H(\hat{d}_A^{\text{opt}})\}$ may then be approximated by the average of the estimated $H(\hat{d}_Q^{\text{opt}})$

and $H(\hat{d}_Q^{\text{opt}})$ across the Monte Carlo data sets, as before.

A.7 Creating “Equivalently Misspecified Pairs” When Both the Propensity Model and Q -function are Misspecified

Consider the $K = 2$ decision point scenario; the developments apply equally to the $K = 1$ setting. To identify pairs $(\beta_{25}^0, \phi_{25}^0)$ that are “equivalently misspecified,” for each of the combinations of β_{25}^0 and ϕ_{25}^0 within a pre-specified grid, say $(\beta_{25}^0, \phi_{25}^0) \in [-1, 1] \times [-1, 1]$ with a step size of 0.05, we generate a large data set of size $n = 10,000$ from the generative data model in Section 5.2 with all other parameters fixed at their true values. This yields $41 \times 41 = 1681$ combinations and hence such data sets. For each data set, the linear regression model for the response and the logistic model for propensity of treatment assignment are then fitted, and the ratio of standard errors for $\hat{\phi}_{25}$ and $\hat{\beta}_{25}$, $SE(\hat{\phi}_{25})/SE(\hat{\beta}_{25})$, say, obtained. We then fit to these values a polynomial model in $\phi_{25}^0, f(\phi_{25}^0)$, say, and select the polynomial degree yielding a sufficiently large adjusted R^2 . Setting $\beta_{25}^0 = \phi_{25}^0/f(\phi_{25}^0)$ then yields the result that the corresponding t-statistics will be approximately equal. These were re-checked in the course of running the simulations so that the t-statistics differed by less than some reasonable value, usually at most a 5 percent difference, as it cannot be guaranteed that they will be precisely the same.

A.8 Derivation of $h_1^0(s_1; \beta_1^0)$ and $C_1^0(s_1; \psi_1^0)$ in the Two Decision Point Scenario

We seek to identify the true $h_1^0(s_1)$ and $C_1^0(s_1)$, where S_1 and A_1 are Bernoulli. With $h_1^0(s_1) = \beta_{10}^0 + \beta_{11}^0 s_1$ and $C_1^0(s_1) = \psi_{10}^0 + \psi_{11}^0 s_1$, it follows that the true Q -function at the first decision is $Q_1^0(s_1, a_1) = h_1^0(s_1) + a_1 C_1^0(s_1)$. We thus calculate $Q_1^0(s_1, a_1)$ under the generative model and equate terms to determine the form of $\beta_{10}^0, \beta_{11}^0, \psi_{10}^0$, and ψ_{11}^0 . The true value function at the second decision is $V_2^0(S_1, S_2, A_1) = h_2^0(S_1, S_2, A_1) + C_2^0(S_1, S_2, A_1)I\{C_2^0(S_1, S_2, A_1) > 0\}$. Thus, $Q_1^0(s_1, a_1) = E\{V_2^0(S_1, S_2, A_1)|S_1 = s_1, A_1 = a_1\} = \beta_{20}^0 + \beta_{21}^0 s_1 + \beta_{22}^0 a_1 + \beta_{23}^0 s_1 a_1 + \beta_{24}^0 E\{S_2|S_1 = s_1, A_1 = a_1\} + \beta_{25}^0 E\{S_2^2|S_1 = s_1, A_1 = a_1\} + E\{C_2^0(S_1, S_2, A_1)I\{C_2^0(S_1, S_2, A_1) > 0\}|S_1 = s_1, A_1 = a_1\}$. The conditional expectations in this expression may be calculated in a manner analogous to that in Section A.5 to obtain the form of $Q_1^0(s_1, a_1)$. It follows that $Q_1^0(0, 0) = \beta_{10}^0$, $Q_1^0(1, 0) = \beta_{10}^0 + \beta_{11}^0$, $Q_1^0(0, 1) = \beta_{10}^0 + \psi_{10}^0$, and $Q_1^0(1, 1) = \beta_{10}^0 + \beta_{11}^0 + \psi_{10}^0 + \psi_{11}^0$, which

may be solved to yield expressions for β_{10}^0 , β_{11}^0 , ψ_{10}^0 , and ψ_{11}^0 .

Acknowledgments

This work was supported by NIH grants R37 AI031789, R01 CA051962, R01 CA085848, P01 CA142538, and T32 HL079896.

References

- ALMIRALL, D., TEN HAVE, T. and MURPHY, S. A. (2010). Structural nested mean models for assessing time-varying effect moderation. *Biometrics* **66** 131–139.
- BATHER, J. (2000). *Decision Theory: an Introduction to Dynamic Programming and Sequential Decisions*. Chichester: Wiley.
- BLATT, D., MURPHY, S. A. and ZHU, J. (2004). A-learning for approximate planning. *Technical Report 04-63, The Methodology Center, Pennsylvania State University*.
- CHAKRABORTY, B., MURPHY, S. A. and STRECHER, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research* **19** 317–343.
- CRAVEN, M. W. AND SHAVLIK, J. W. (1996). Extracting tree-structured representations of trained networks. In *Advances in Neural Information Processing Systems*, volume 8, 24–30. Denver, CO: MIT Press.
- HENDERSON, R., ANSELL, P. and ALSHIBANI, D. (2010). Regret-regression for optimal dynamic treatment regimes. *Biometrics* **66** 1192–1201.
- LAVORI, P. W. and DAWSON, R. (2000). A design for testing clinical strategies: biased adaptive within-subject randomization. *Journal of the Royal Statistical Society, Series A* **163** 29–38.
- LABER, E. B. and MURPHY, S. A. (2011). Adaptive confidence intervals for the test error in classification. *J. Amer. Statist. Assoc.* **106** 904–913.

- LABER, E. B., QIAN, M., LIZOTTE, D. J. and MURPHY, S. A. (2010). Statistical inference in dynamic treatment regimes. Pre-print, arXiv:1006.5831v1.
- MOODIE, E. E. M., RICHARDSON, T. S. and STEPHENS, D. A. (2007). Demystifying optimal dynamic treatment regimes. *Biometrics* **63** 447–455.
- MURPHY, S. A. (2003). Optimal dynamic treatment regimes (with discussion). *J. Royal Statist. Soc., Ser. B* **58** 331–366.
- MURPHY, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Stat. Med.* **24** 1455–1481.
- MURPHY, S. A., LYNCH, K. G., OSLIN, D. MCKAY, J. R., and TEN HAVE, T. (2007a). Developing adaptive treatment strategies in substance abuse research. *Drug Alcohol Depend.* **88S** S24–S30.
- MURPHY, S. A., OSLIN, D. W., RUSH, A. J. and ZHU, J. (2007b). Methodological challenges in constructing effective treatment sequences for chronic psychiatric disorders. *Neuropsychopharmacology* **32** 257–262.
- NAHUM-SHANI, I., QIAN, M., ALMIRALL, D., PELHAM, W. E., GNAGY, B., FABIANO, G., WAXMONSKY, J., YU, J. and MURPHY, S. A. (2010). Q-Learning: A data analysis method for constructing adaptive interventions. Technical report.
- ORELLANA, L., ROTNITZKY, A and ROBINS, J. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part I: Main content. *Int. J. Biostatist.* 6, Issue 2, Article 8, DOI: 10.2202/1557-4679.1200.
- ROBINS, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods: Applications to control of the healthy worker survivor effect. *Math. Model.* **7** 1393–1512.
- ROBINS, J. M. (1994). Correcting for non-compliance in randomized trials using structural nested mean models. *Comm. Statist. – Theory Meth.* **23** 2379–2412.
- ROBINS, J. M. (2004). Optimal structured nested models for optimal sequential decisions. In *Proceedings*

- of the Second Seattle Symposium on Biostatistics, D. Y. Lin and P. J. Heagerty (eds), 189–326. New York: Springer.
- ROBINS, J., ORELLANA, L. and ROTNITZKY, A. (2008). Estimation and extrapolation of optimal treatment and testing strategies. *Stat. Med.* **27**, 4678–4721.
- ROSTHØJ, S., FULLWOOD, C., HENDERSON, R. and STEWART, S. (2006). Estimation of optimal dynamic anticoagulation regimes from observational data: A regret-based approach. *Stat. Med.* **25** 4197–4215.
- RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.
- RUSH, A. J., FAVA, M., WISNIEWSKI, S. R., LAVORI, P.W., TRIVEDI, M.H., SACKeim, H. A., THASE, M. E., NIERENBERG, A. A., QUITKIN, F.M., KASHNER, T. M., KUPFER, D. J., ROSENBAUM, J. F., ALPERT, J., STEWART, J. W., MCGRATH, P. J., BIGGS, M. M., SHORES-WILSON, K., LEBOWITZ, B. D., RITZ, L., NIEDEREHE, G. (2004). Sequenced Treatment Alternatives to Relieve Depression (STAR*D): rationale and design. *Control. Clin. Trials* **25** 119–142.
- RUSH, A. J., TRIVEDI, M.H., IBRAHIM, H.M., CARMODY, T.J., ARNOW, B., KLEIN, D.N., MARKOWITZ, J.C., NINAN, P.T., KORNSTEIN, S., MANBER, R., THASE, M.E., KOCSIS, J.H., AND KELLER, M.B. (2003). The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry* **54**, 573–583.
- SHORTREED, S. M., LABER, E., LIZOTTE, D. J., STROUP, T. S., PINEAU, J. and MURPHY, S. A. (2010). Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Mach. Learn.* **11** 109–136.
- SONG, R., WANG, W., ZENG, D., and KOSOROK, M. R. (2010). Penalized q-learning for dynamic treatment regimes. Pre-Print, arXiv:1108.5338v1.
- THALL, P. F., MILLIKAN, R. E. and SUNG, H. (2000). Evaluating multiple treatment courses in clinical trials. *Stat. Med.* **19** 1011–1028.

- THALL, P. F., SUNG, H. and ETSEY, E. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *J. Amer. Statist. Assoc.* **97** 29–39.
- THALL, P. F., WOOTEN, L. H., LOGOTHETIS, C. J., MILLIKAN, R. E., and TANNIR, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Stat. Med.* **26** 4687–4702.
- WATKINS, C. J. C. H. (1989). Learning from Delayed Rewards. Ph.D. Thesis. King’s College, Cambridge, U.K.
- WATKINS, C. J. C. H. and DAYAN, P. (1992). Q-learning. *Mach. Learn.* **8** 279–292.
- ZHAO, Y., KOSOROK, M. R. and ZENG, D. (2009). Reinforcement learning design for cancer clinical trials. *Stat. Med.* **28** 3294–3315.